ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 39, 98–114 (1987)

Facing Uncertainty in the Game of Bridge: A Calibration Study

GIDEON KEREN

Institute for Perception TNO, Kampweg, The Netherlands

Previous studies have shown a strong tendency toward overconfidence in peoples' probability assessments. Even experts are often poorly calibrated. The present paper suggets that quality of calibration is largely determined by the extent to which the cognitive processes required for repeated probability assessments are similar. One task that satisfies this condition is the game of bridge in which accurate probabilistic assessments are required for good playing. Two experiments were conducted in a natural setting of a tournament in which subjects were asked to assess the likelihood that a final contract (reached during the bidding phase) would indeed be made. Expert players were almost perfectly calibrated whereas amateurs were overconfident. The differences between expert and amateur players are discussed, and some guidelines for training procedures for calibration in general are proposed.

A robust finding in the decision literature concerns the so-called overconfidence and miscalibration phenomena. A judge is said to be well calibrated if "over the long run, for all propositions assigned the same probability, the proportion true is equal to the probability assigned" (Lichtenstein & Fischhoff, 1977, p. 161). Much of the research on calibration shows that people are poorly calibrated (Lichtenstein, Fischhoff, & Phillips, 1982); that is, their subjective probability assessments deviate considerably from the true probabilities. In particular, they most often express overconfidence by assigning probabilities higher than warranted (e.g., Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977).

What is implied when we say that a person is poorly calibrated? According to the definition of Lichtenstein and Fischhoff (1977), calibration is measured in the *long run* and as such has a "frequentistic" interpretation (see also Lindley, 1982) that is based on repeated events.

Previous researchers, especially those in the Bayesian tradition, have viewed subjective probabilities as a property of the judge. Notwith-

I thank Carol Varey for her perceptive comments on various stages of this project. Reprints may be obtained from Gideon Keren, Institute for Perception TNO, Kampweg 5, 3769 DE Soesterberg, The Netherlands.

0749-5978/87 $3.00

standing that statement, the extent to which a person is well calibrated may depend, among other things, also on the nature of the stimulus material that is being judged. In particular, given a set of repeated events or items, one can distinguish between what may be called *related* and *unrelated* items. For example, much of the research on calibration showing overconfidence has employed general knowledge questions like "What is the population of Peru?" "What is the capital of Nepal?" or "What is the longest river in the world?" For this type of stimulus material the subjects' knowledge of one item is independent of their knowledge of another item. When no information (knowledge) can be inferred from one item and transferred to another one we deal with unrelated items.

When calibration of probabilities is involved, the question is to what extent are items in a given set sufficiently related so that knowledge about the occurrence (or nonoccurrence) of some items can be used for assessing the probability of other items in the same set. It is important to emphasize that related items do not have to be dependent. For instance, consider a bridge player who has been exposed to thousands of bridge hands. The different hands are strictly independent and yet are related items.

The judgment of the extent to which items are related, in the above sense, is a subjective one and refers to a great extent to the similarity in the subject's mental processes when reacting to one item or the other. In certain tasks the items are clearly related in that they require similar repetitive cognitive processes to be applied. Under such conditions learning may take place and "good" calibration can be achieved. Weather forecasting may serve as a good example for such a task. When meteorologists state that there is a 60% chance of rain they actually mean to say that 60% out of all previous cases (items) with similar weather conditions (as reflected in barometer readings, wind directions, satellite pictures, etc.) resulted in rain. Weather forecasts form related items and thus meteorologists can benefit from their previous experience and be well calibrated.

I propose to distinguish between different types of calibration tasks depending on the extent to which items are related (as explained above) and require similar cognitive processes. Unrelated items require different cognitive processes for each item so that feedback is ineffective in calibrating those processes. When, however, items in a task are related, there are grounds to develop inferential processes that would lead to probabilistic notions in terms of relative frequency and long run considerations. I suggest that such conditions are necessary (though not sufficient) for subjects to be well calibrated. Note that the notion of related items as employed here is subjective and is measured on a continuous

scale. The higher the degree of relatedness, the better the grounds for developing good calibration.

To obtain good calibration when items are related, two additional requirements have to be satisfied: Feedback has to be provided, and sufficient practice has to take place. Feedback to unrelated items is of little use, since it carries no information with regard to other items (events), and hence does not enable one to develop inferential processes and a prediction model. Feedback to related items, in contrast, provides information that in the long run may, if applied correctly, be used for assessment of the relevant variables and consequently to revision of probabilities of future events. The excellent calibration performance by weather forecasters (Daan & Murphy, 1982; Murphy, 1981; Murphy & Winkler, 1984) can probably be accounted for by, among other things, the prompt and continuous feedback as well as by long enduring practice.

The role of practice and training is not yet clear, and the results of calibration studies with experts are quite varied. For instance, Lichtenstein and Fischhoff (1977) asked graduate students in psychology to respond to 50 two-alternative general-knowledge items and 50 items covering knowledge in psychology (in which they were supposedly experts). The two subsets were of equal difficulty, and calibration was poor and similar for the two tasks. Also, several studies (e.g., Christensen-Szalanski & Bushyhead, 1981; Lusted, 1977) have shown physicians' probability assessments of their diagnosis to yield poor calibration. Among studies on experts, only weather forecasters exhibit superb calibration. These same meteorologists, however, are not better calibrated than student subjects when the task is general knowledge questions (Keren, 1985). This suggests that calibration by experts may be task specific. To assess, however, the role of expertise in calibration one needs a control group of nonprofessionals. Most of the calibration experiments on experts lack such a control group.

To assess the contribution of expertise I have chosen to study bridge players with different levels of expertise. The advantage of choosing such a task is that the difference in expertise arises in a natural way rather than being manipulated in the laboratory. The study was conducted in a natural setting, a tournament, with highly motivated subjects. Using the game of bridge has additional advantages: The items, namely the different games played by the subjects (both during the experiment as well as in their previous history), are related items with continuous feedback provided after each game; hence, we would expect bridge players to be well calibrated. This expectation is also based on the fact that probability assessments are a natural and integral part of the game. In other words, good calibration is a necessary requirement for a good bridge player.

Previous research (Lichtenstein *et al.*, 1982) suggests that quality of calibration, and in particular overconfidence, are strongly related to task difficulty. As tasks get easier, overconfidence is reduced. One may reverse the question and ask whether for a given task where difficulty (in terms of task requirements) remains constant, does improvement in skill (i.e., expertise) also imply improved calibration? The present study makes an attempt to answer this question.

A detailed description of the game of bridge is beyond the scope of the present paper (see Epstein, 1967; Goren, 1952; Kaplan, 1963). It will be sufficient to mention here a few aspects of the game that are relevant for the present study. Briefly, the game consists of two parts: The bidding and the play. The bidding is a communication process which occurs under certain sequential restraints and with a limited vocabulary, and terminates with a final contract. The final contract is a bet, by one of the two teams, to take a certain number of tricks (with a particular suit as trumps) during the second phase of the game called the play. The goal of the team that declared the final contract (the offense) is to make as many tricks as promised or even more. The other team (the defense) is trying to prevent the fulfillment of the contract. After the bidding is over, one of the defenders leads a card, then all the cards of one of the players in the offense are laid down on the table and become the dummy. Then the play follows, each team trying to take as many tricks as possible.

The bidding phase is aimed at reducing the uncertainty but can never eliminate it completely. At the end of the bidding the declarer (the player of the offensive team who is actually playing) faces two uncertainties: The partner's exact card combination is not yet known, nor is that of the defense. The second type of uncertainty concerns the manner in which the defense team is going to play. Obviously, the final outcome of the play is determined by the way in which both teams play. Given these uncertainties, players have to make accurate probability assessments of making a certain contract, before they bid. Thus, good bridge players have to be well calibrated. The following experiments were designed to test this hypothesis.

EXPERIMENT 1

Method

Subjects. The subjects were 16 highly experienced players (8 pairs) from one of the top bridge clubs in The Netherlands. All have participated in national tournaments and several of them have participated in international competitions as well.

Procedure and stimuli. The experiment was conducted during an evening tournament organized by the experimenter. There were four tables

(two pairs for each table) and 28 decks of cards, each constituting a game. Each deck was divided in advance into four "hands" of 13 cards each, and remained the same for the entire tournament. The 28 decks were divided into seven rounds of four games. Each pair played a round (four games) against each of the other seven pairs according to a predetermined order. Thus, each of the 16 players played 28 games during the tournament (to total 448 observations) and each game was played four times. Financial prizes were awarded to the first three pairs. The rest of the players received a fixed amount of 20 Dutch guilders each (approximately $7).

Before the tournament started the subjects were given the following instructions: They were told that at the end of the bidding and before the play started (i.e., also before the dummy cards were laid down) they were to estimate the probability that the final contract would be made. Probability statements were made by using numbers between 0 and 100 (percentages). Subjects were instructed that 100 meant they were absolutely sure that the contract would be made. Similarly, 0 meant that they were absolutely sure that the contract would fail. Further, they could use any number between 0 and 100 to indicate their confidence. Low ratings meant that the contract was likely to fail and high ratings meant they thought the contract would be made. A 50% rating meant they believed there was an equal chance of success or failure of the contract. Each player received a sheet with the games numbered 1 to 28 and was asked to provide a probability assessment after the bidding phase of each game. Players were required to make these assessments individually and refrain from any exchange of information except that allowed by the bidding rules. All four players made an independent assessment and also noted the game number, their position (whether they played in the north, east, south, or west position), and their role (offense or defense). After all players had completed this task, they went on to the playing phase. The final score was recorded on a separate sheet.

Results and Discussion

Out of 112 games played, 63 (56%) ended up with contracts that were made.[1] One way to assess how well players were calibrated is via a calibration curve, which is a graph showing the hit rate (percentage correct) for each probability response. Calibration curves were constructed by

[1] In general, one should not use the percentage of fulfilled contracts as a criterion for players' quality. Expert players often bid intentionally a contract they know cannot be made, with the hope that they will lose less points than the other team may gain, if the other team will make the final bid. In addition, a low percentage of made contracts can also indicate outstanding defense.

grouping (over subjects and games) all the responses into 12 categories, in the ranges 0, .01–.10, .11–.20, .21–.30 . . . , .70–.79, .80–.89., .90–.99, and 1.00.

The resulting calibration curve is shown in Fig. 1. A calibration curve by itself may not be an adequte presentation, because it does not take into account the relative weights, in terms of number of observations of each point. In particular, some points on the curve are based on a very small number of observations and hence constitute unreliable estimates. An attempt was therefore made to fit a model to the data. Given the inherent boundaries of 0 and 1.0 for confidence ratings, it is reasonable to consider a model that reflects these constraints. An approach that satisfies these contraints, and which is adopted here, is to work with *logistic* linear models.

Assume the frequencies of correct predictions f_c for each confidence rating category c ($c = 0, 1, 2, 3 \ldots 11$) to be independent binomial random variables with probabilities of being correct π_c, mean confidence rating for category c to be r_c, and number of observations n_c. The probability of being correct can then be described as

$$\pi_c = \frac{\exp(\alpha + \beta r_c)}{1 + \exp(\alpha + \beta r_c)} \tag{1}$$

where α and β play roles that are similar to "intercept" and "slope,"

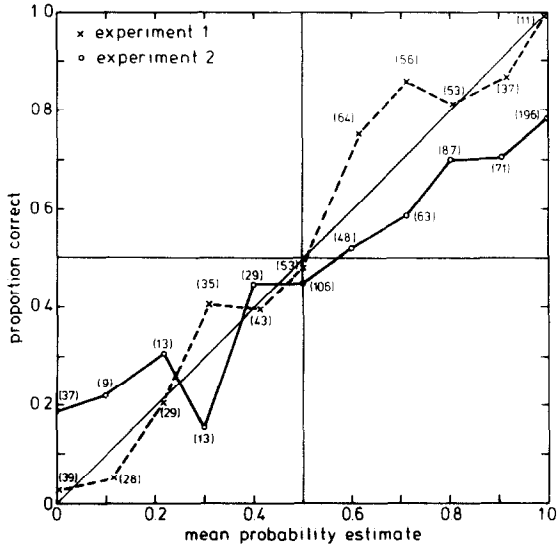

FIG. 1. Calibration curves for expert (Experiment 1) and amateur (Experiment 2) players (numbers in parentheses indicate number of observations).

respectively. Expression (1) can also be written as a simpler linear model for the log-odds or logit of π_c

$$\log_e \left[\frac{\pi_c}{1 - \pi_c} \right] = \alpha + \beta r_c \tag{2}$$

In other words, the proportion of correct predictions of the outcome of a game is expressed as a function of the confidence ratings. The calibration curve based on such a model is presented in Fig. 2. It should be emphasized that the models employed in the present paper are used for purposes of data analysis and no attempt should be made to interpret them otherwise (e.g., as a process model).

All estimates and tests were obtained with the help of the GLIM computer program (Baker & Nelder, 1979), which applies the method of maximum likelihood to generalized linear models (Nelder & Wedderburn, 1972). To evaluate the model, I used the deviance which is a likelihood ratio measure of the difference between the models' predictions and the actual data. Those readers who lack detailed knowledge of generalized linear interactive modeling may simply interpret the models in the more familiar framework of an ANOVA, and treat the deviance as representing the SS_{res} (sum of squares of the residuals) for a model. The deviance for the model was 10.74 which according to a χ^2 test ($df = 9$) is not signifi-

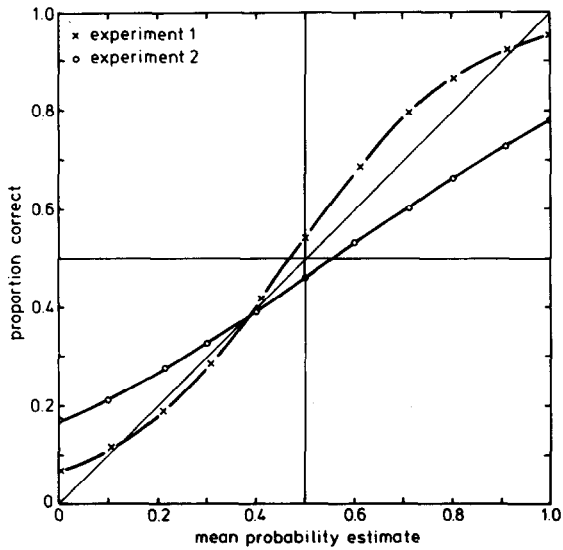

FIG. 2. Calibration curves for expert (Experiment 1) and amateur (Experiment 2) players constructed from the predictions of the logit model.

cant ($p < .25$). I thus conclude that the model is a reliable representation of the data.

Inspection of both Figs. 1 and 2 suggests that the expert players in our sample are indeed very well calibrated. In particular, their confidence ratings almost coincide with the 45° line for the range of 0–.50 confidence ratings, and are slightly underconfident for the range of .50–.100.

The calibration curve shows averages across subjects and may conceal possible individual differences. In Fig. 3 the crosses depict, for each player separately, the mean probability estimate (across the 28 games played) against the corresponding percentage of contracts made. Despite the small number of observations for each player, represented in the figure by a point, it is apparent that most players are well calibrated, that is, most of the points lie relatively close to the 45° line. Taken together, the data suggest that the expert players in our sample were well calibrated. However, to attribute the good calibration to expertise one would need a control group to show that poorer players are less well calibrated. Indeed, most calibration studies on experts have failed to employ such a control group. Before we further analyze the data of the expert players a second experiment is reported, which provides the desired control and serves as a base line to which experts' performance can be compared.

EXPERIMENT 2

Method

Subjects. The subjects were 28 members (14 pairs) of a sport club

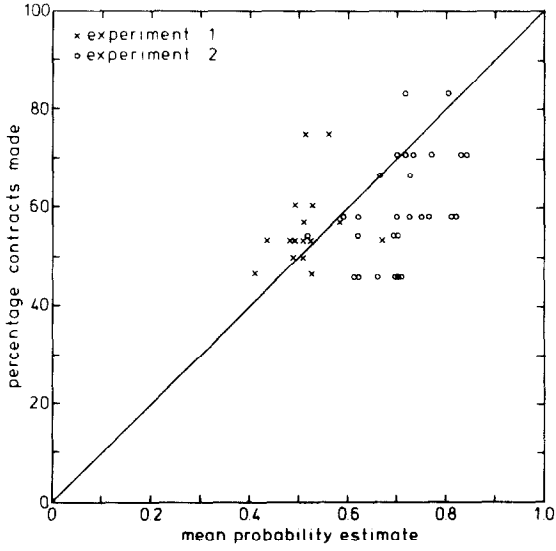

FIG. 3. Mean probability estimate (that the contract would be made) for each individual player plotted against actual percentage of contracts made.

which among other activities also organized a bridge tournament. Although all had been players for a long time, their frequency of playing was much lower than that of subjects in Experiment 1. None of them had ever participated in a national or international competition.

Procedure and stimuli. Procedure and stimuli were identical to Experiment 1 with the following exceptions: There were 14 pairs of players and 24 games (or decks). The 24 games were identical to the first 24 games played in Experiment 1. Since amateur players are slower, the last 4 games of Experiment 1 were omitted, thus reducing the number of rounds in the tournament to six. Each player played all the 24 games (six rounds) and each game was played seven times. Total number of observations was 672 (28×24). In all other respects the procedure was identical to that of Experiment 1.

Results and Discussion

Of the 168 games played, 101 (60%) ended up with contracts that were made. The calibration curve for subjects in Experiment 2 is portrayed in Fig. 1 (circles). The predicted calibration curve resulting from fitting the *logistic* linear model (as in Experiment 1) is shown in Fig. 2. The fit of the model was even better than in Experiment 1, the deviance being equal to 3.37, and the corresponding χ^2 test ($df = 9$) was nonsignificant ($p < .9$). The even smaller deviance of the model for the amateurs data is not surprising. It implies that amateurs conform closely to the nonoptimal calibration curve produced by the logit model. The logit model also offers a good representation of experts' calibration except for the two ends of the curve (confidence ratings of 0 and 1.0) at which experts are almost perfectly calibrated. It is these two extreme points that contribute to a somewhat larger deviance for the experts' data.

As can be seen from Figs. 1 and 2, amateur players in Experiment 2 are poorly calibrated, especially when compared with the expert group. They are overconfident when assigning high confidence ratings and underconfident when assigning low confidence ratings. Inspection of individual players in Fig. 3 (circles) suggests that the large majority of players exhibit overall overconfidence in their probability assessments. One way to estimate the tendency to be overconfident or underconfident is to take the difference between the mean confidence rating and the corresponding percentage of contracts made, for each individual player. Averaging these differences across players we obtain a mean difference of $-.046$ (underconfidence) and .107 (overconfidence) for players in Experiments 1 and 2, respectively. Since positive and negative differences may cancel each other, I also computed the absolute differences which indicate the distance from perfect calibration (i.e., the 45° line). The mean absolute differences across subjects were .072 and .121 for Experiment 1 and 2, re-

spectively. This difference between the two groups of players was significant ($t(42) = 2.13$, $p < .025$).

A major source of the difference between expert (Experiment 1) and amateur (Experiment 2) players lies in the frequency with which they use extreme confidence ratings. Table 1 shows the frequency distribution of confidence ratings for both Experiments 1 and 2. Confidence ratings of 1.0 account for more than 29% of amateurs' assessments and only for 2.5% for experts. Moreover, for all the games in which experts assigned a rating of 1.0 the contract was always made, thus yielding a hit rate of 1.0. In contrast, the corresponding hit rate for amateurs was only .78, that is in 22% of the games in which a confidence rating of 1.0 was assigned the contract actually failed. Both experts and amateurs use a 0 confidence rating (i.e., probability 1.0 that the contract will fail) approximately equally often. However, the corresponding hit rates in this case are .97 and .81 for experts and amateurs, respectively. In summary, extreme confidence ratings are not only used by experts less frequently, they are also used more appropriately.

One may be tempted to explain the above results by simply stating that experts are more conservative in their estimates. This explanation, however, is untenable for two reasons; First, as already shown, the accuracy of experts is much better than that of amateurs on extreme ratings. A second argument that refutes the "conservatism" hypothesis is provided from another independent observation of the game. The vocabulary of bidding in bridge contains an option of *doubling:* If a player believes that the most recent bid of the opponents is unmakable he or she may double that contract. A "double" announcement is equivalent to saying that the player believes the probability of the contract being made to be very low indeed.[2] Doubles have benefits and costs: The benefit is that if the contract indeed fails, the defenders who doubled receive double the number

TABLE 1
FREQUENCY DISTRIBUTION OF CONFIDENCE RATINGS FOR EXPERT (EXPERIMENT 1) AND AMATEUR (EXPERIMENT 2) PLAYERS

Confidence rating	Experiment 1		Experiment 2	
0	33	(7.4%)	37	(5.5%)
.01–.33	91	(20.3%)	35	(5.2%)
.34–.66	167	(37.3%)	183	(27.2%)
.67–.99	146	(32.5%)	221	(32.9%)
1.00	11	(2.5%)	196	(29.2%)
Total	448	(100%)	672	(100%)

[2] The doubles under discussion are only doubles on the final contract (i.e., do not include "information" doubles).

of points they would otherwise get. However, if the contract is nevertheless made the offenders will gain more points than they would otherwise get. Hence, books on bridge usually advise players to use the double option with caution. Judging by the frequency of using the doubling option, expert players were clearly less conservative: They used the option on 26 out of 112 games (23.2%) compared with the amateurs 17 out of 168 games (10.1%). More important, only in one single case (4%) was the contract made despite the double announcement in the expert tournament, compared to 8 games (47%) in the case of amateurs. Thus, experts used the doubling option more frequently and more accurately.

There is another analysis that points out the differences between expert and amateur players which may suggest some differences in the underlying cognitive processes of the two groups. Consider those games in which the player is relatively sure about the outcome. For instance, consider all the ratings of .80 and higher, which suggest that the player is quite confident that the contract will be made, together with all the ratings of .20 or less, in which the player is quite confident that the contract will fail. An additional classification can be made by separating ratings that were made by players from the offensive team from those made by the defensive team. Table 2 shows the frequency of observations in each of the four cells of a 2 × 2 table framed by ratings (above .80/below .20) and role (offense/defense) for Experiments 1 and 2, respectively. A χ^2 analysis on the data from Experiment 1 yields no significant effects. However, for Experiment 2 there is a significant main effect of ratings ($\chi^2 = 215$, $p < .001$) and a significant interaction ($\chi^2 = 5.17$, $p < .025$).[3]

The interaction is in particular revealing since it shows that amateurs tend to assign low probabilities more often when they are defending, that is, believing (or hoping) that the opponents will fail. In contrast, they assign a high probability of success to their own contracts more frequently than to their opponents. The lack of significant effects in Experiment 1 suggests that experts are able to assess probabilities more objectively without contaminating it with their own wishes.

The choice of using confidence ratings of .80 and .20 as cutoff points for above .79 and below .21 as representing high confidence is somewhat arbitrary. However, an identical analysis performed on ratings of .90 and above and .10 and below led to the same pattern and identical conclusions.

Forecasting is intimately tied to decision making and should be evalu-

[3] Strictly speaking, such an analysis should be carried out for each individual player. Unfortunately, this was impossible due to the small number of observations (hands) for each player. However, an informal check suggests that the overall pattern holds also for the large majority of the individual players.

TABLE 2
FREQUENCY OF HIGH AND LOW CONFIDENCE RATINGS AS A FUNCTION OF ROLE OF PLAYER AND EXPERTISE

	Experiment 1		Experiment 2	
Confidence ratings	Offense	Defense	Offense	Defense
.20 or less	50	39	20	35
.80 or more	54	47	185	165

ated within this context (Einhorn & Hogarth, 1982). Unfortunately, in the large majority of calibration studies, forecasting and decision making have been separated and the link between the two has been left vague. Since, as was argued earlier, forecasting the odds in bridge is a natural and integral part of the decision process, it is possible here to link these two aspects together. In particular, we may assume that any confidence rating less than .50 implies that it is more likely that the contract will fail. A confidence rating higher than .50 represents the belief that it is more likely that the contract will be made. Excluding all confidence ratings of .50, we analyzed the relative frequency that contracts will be fulfilled or fail, as a function of whether the corresponding confidence ratings are above or below .50. These frequencies are given in Table 3. As can be seen, in each of the 2 × 2 tables (corresponding to each experiment) there is a clear first-order interaction suggesting that there is a high relationship between a confidence rating being above or below .50 and the likelihood that the contract will finally be made (or fail). However, this relationship is not perfect. Players sometimes assign confidence ratings less than .50 to games that are actually made, and confidence ratings higher than .50 to games that actually fail. The rate of such errors is much higher for the amateur subjects in Experiment 2. Using an iterative computation of the expected values (Fienberg, 1977), we obtained a significant second-order interaction of the 2 × 2 × 2 contingency analysis (confidence above or below .50 contract made or failed, Experiment 1 or 2) with $\chi^2_1 = 9.82$ which is highly significant, $p < .005$.

TABLE 3
FREQUENCY OF MAKING (FAILING) A CONTRACT AS A FUNCTION OF CONFIDENCE RATINGS HIGHER OR LOWER THAN 50% FOR EACH OF THE TWO EXPERIMENTS

	Experiment 1			Experiment 2		
	Confidence below .50	Confidence above .50		Confidence below .50	Confidence above .50	
Contracts made	40	188	226	30	326	356
Contracts failed	132	37	169	71	140	211
Total	172	225		101	466	

The marginals in Table 3 shed additional light on the differences between experts and amateurs. Note that for experts (Experiment 1) there is an almost perfect match between the relative frequency of using confidence ratings above or below .50 and the corresponding proportions of contracts made or failed, which can actually be considered as the base rate. For the amateurs (Experiment 2) the proportion of ratings above .50 is much higher than the corresponding proportion of contracts made, suggesting a bias of "optimism" (Weinstein, 1980) congruent with the overconfidence in the calibration curves.

GENERAL DISCUSSION

The game of bridge requires decisions that have to be made in the face of uncertainty. In fact, at least two different sources of uncertainty are inherent in the game: One stems from imperfect knowledge of the card distribution. The second source of uncertainty (not necessarily independent from the first one) is that players cannot know for sure how their opponents will play. More generally, they cannot judge in advance whether their own decisions or their opponents' decisions are optimal; this is only known when the game has ended and all the cards are open. Different decisions by either the player or the opponents may often lead to different outcomes, and these have to be taken into account during the bidding phase.

The results of the two studies reported here suggest that amateur players are inferior to expert players in taking both kinds of uncertainties into account during the game. As far as card distribution is concerned, the hit rate for 0 confidence ratings by amateur players is indicative. A confidence rating of 0 should imply that the player is absolutely confident that the contract will fail *independent* of how the offense plays it. Nevertheless, amateur players erred in 19% of the cases in which they assigned a confidence rating of 0, compared to 3% errors made by experts. Further support for this observation is provided by the relatively low accuracy of amateurs when using the doubling option. A double should reflect the player's belief that the card distribution is such that the contract is extremely unlikely to be made, independently of how it is played. However, in more than 47% of the cases in which amateurs used the option of doubling the contract was nevertheless made. For experts, in contrast, the corresponding failures with doubling were a meager 2%.

The other dimension of uncertainty, which amateurs often fail to appreciate, is due to lack of knowledge of how the playing phase (after the bidding) will develop. There are several possible decisions on each move of either the offense or the defense, so these cannot be predicted with certainty. Amateurs, however, assume implicitly that the play will de-

velop according to some optimal considerations (or at least what they consider optimal), and they do not leave room for variability (and hence larger uncertainty) of outcomes due to imperfections of players. This conjecture is supported by two related observations; one is the amateurs' disproportionately frequent use of high compared to low confidence ratings (no significant difference for experts). The second observation concerns confidence ratings of 1.0. A close check of all the contracts to which amateurs assigned a 1.0 confidence suggests that as far as the card distribution is concerned the contract could theoretically have been made in almost all the cases given an optimal play (such a theoretical analysis can obviously be made only with perfect card certainty, i.e., when all cards are laid open). Nevertheless, in 22% of those cases the contract failed. Since all these contracts were theoretically possible, the failure is probably due to nonoptimal playing, a dimension which as we suggest is not sufficiently taken into account by amateurs. None of the games that received ratings of 1.0 by experts failed.

An alternative explanation (that does not exclude the previous one) for the amateurs' failure to properly evaluate the playing phase, and that can also account for their overconfidence, may be due to misconception of their own skill. Specifically, amateurs may know that they are error prone, but operate under the assumption that they will make fewer mistakes than their opponents. A similar observation has been reported by Svenson (1981) who demonstrated that the majority of subjects in his sample regarded themselves as more skillful drivers compared with other subjects in the sample. Support for the above explanation is obtained in the present study from the observation concerning the difference in assigning confidence ratings depending on whether a player takes the offensive or defensive role. For expert players there is no difference in the pattern of ratings depending on role. Amateurs, in contrast, show a significant tendency to assign higher ratings to contracts they are trying to make and lower ratings to contracts they are trying to defeat. Thus, amateurs exhibit a bias of "optimism" (Weinstein, 1980), whereas experts are able to make more objective judgments (that are detached from their motives and goals).

The above observations can account for the poor calibration of the amateur players, in particular the strong overconfidence shown in Fig. 1. At the same time, the exceptionally "good" performance of expert players suggests that good calibration in this task is possible. What requirements are needed to achieve this?

To account for the experts "good" calibration, and the relatively poor performance of amateur players, consider the following hypotheses.[4] As-

[4] Acknowledgments are made to C. Gettys for his invaluable contribution to this part of the discussion.

sume that the process of generating confidence ratings, or probability assessments, is composed of two subprocesses.

One is the process of (semantic) inference where a person builds a mental model of a situation based on a knowledge base (usually derived from own experience), and uses this model to generate nonnumerical feelings of certainty (e.g., Beyth-Marom, 1982). For instance, these feelings can be expressed as the plausibility of various conflicting scenarios that lead to different outcomes. The nature or quality of this process is mainly determined by two factors: One concerns the extent to which the initial mental model that is constructed is the proper one, and similarly that the appropriate inferential processes are employed. By a proper mental model I mean that a correct suitable "problem space" (Keren, 1984; Newell & Simon, 1972), for a given situation, is adopted. The second factor affecting this process concerns the amount of data, based on experience, that is fed into the mental model.

The second subprocess is one in which these feelings of plausibility and uncertainty are translated into numerical estimates, that is, into probabilities. Let us further assume that for a subject to be "well calibrated" both processes must be "well tuned."

The results of the two studies reported in this article may now be interpreted in light of the proposed framework. Specifically, it was suggested in the introduction that for tasks with highly related items (as is the case in bridge), a sufficient amount of practice and feedback may eventually lead to good calibration.

The critical aspect of relatedness of items is how a person learns from feedback. If items are unrelated, as is the case with calibration studies using general knowledge questions, then feedback cannot improve the first subprocess. A subject faced with a task composed of unrelated items may still learn a functional transformation used in the second subprocess, but this will lead only to very limited improvement and will not generalize to any other task. For instance, Lichtenstein and Fischhoff (1980) report two experiments, using general knowledge questions, in which the quality of people's probability assessments improved through intensive training. However, as these investigators showed, there was very little if any generalization to several related probability assessment tasks. Moreover, the improvement they obtained is technical in nature. Obviously, subjects who receive continuous feedback may soon find out that their probability assessments are too high or too low and change their assessments accordingly. I suggest, however, that such a change reflects at best a modification of the translation rule in the second subprocess, but leaves the first (and more important) subprocess unchanged.

When items are related, the tuning of both subprocesses can (but do not necessarily have to) be improved by feedback. Physicians, for instance, and specialists in particular, deal with related items and suppos-

edly learn from feedback to elaborate their mental models and improve their insights. Indeed, with experience they usually become better physicians. The reason that these physicians are nevertheless poorly calibrated (e.g., Lichtenstein *et al.*, 1982) is explained by the fact that ordinarily making numerical probability estimates is not part of their job. Consequently their second subprocess remains untuned and this leads to poor calibration. Weather forecasters are one of the few groups that deal with related items, where both subprocesses are becoming well tuned as part of their job.

The expert bridge players in Experiment 1 resemble in many respects the weather forecasters. They deal with highly related items, they have a lot of experience and practice with prompt and immediate feedback, and can thus tune successfully the two necessary subprocesses. Consequently, they exhibit good calibration similar to that shown in meteorologists.

Why was calibration of amateurs, despite their similar long experience, so different from that of the experts group? I propose that the reason is to be found mainly in the first subprocess. Bridge is an extremely complex game and, like chess (De Groot, 1965, 1966), can be played at different levels of expertise. These differences in expertise are reflected in different perceptual strategies and different ways of processing information. Experts differ from amateurs not necessarily in the amount of practice per se, but rather by constructing an appropriate structural representation of the problem that also taps the relevant variables (and their relative weights). Thus, expert bridge players employ, in the first subprocess, an inferential procedure that is better tuned, more sophisticated, and more sensitive to details as compared with amateur players.

The above conclusions have important implications for training procedures for calibration (Lichtenstein & Fischhoff, 1980). They suggest that a general training method intended to improve calibration independently of the task is most likely to fail. A successful training program should be task-specific, and ensure that the training would lead to a clear and deep understanding of the structure and variables that are involved in the inferential subprocess, and at the same time provide sufficient cues and practice for the second subprocess, namely the translation of knowledge into numerical probabilities. A procedure that takes those considerations into account should eventually lead to a calibration curve similar to the one produced by expert bridge players.

REFERENCES

Baker, R. J., & Nelder, J. A. (1979). *The GLIM system generalized linear interactive modeling* (Release 3). England: Rothamsted Experimental Station, Herpenden, Herts.

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

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 928–935.

Daan, H., & Murphy, A. H. (1982). Subjective probability forecasting in The Netherlands: Some operational and experimental results. *Meteorol. Rdsch.* 35, 99–112.

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

De Groot, A. D. (1966). Perception and memory versus thinking. In B. Kleinmuntz (Ed.), *Problem solving*. New York: Wiley.

Epstein, R. A. (1967). *The theory of gambling and statistical logic*. New York: Academic Press.

Einhorn, H. J., & Hogarth, R. M. (1982). Prediction, diagnosis, and causal thinking in forecasting. *Journal of Forecasting* 1, 23–36.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press.

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.

Fischhoff, B., & Slovic, P. (1980). A little learning . . . : Confidence in multicue judgement. In R. Nickerson (Ed.), *Attention and performance* (Vol. 8). Hillsdale, NJ: Erlbaum.

Goren, C. H. (1952). *Contract bridge complete*. Garden City, NY: Doubleday.

Keren, G. (1984). On the importance of identifying the correct "problem space." *Cognition*, 16, 121–128.

Keren, G. (1985). *On the calibration of experts and lay people*. Unpublished manuscript.

Kaplan, E. (1963) *Winning Contract Bridge Complete*. N.Y.: Bentam Books.

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

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 of 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge Univ. Press.

Lindly, D. V. (1982). The improvement of probability judgements. *Journal of the Royal Statistical Society*, A, 145, P. 1, 117–126.

Lusted, L. B. (1977). *A study of the efficacy of diagnostic radiologic procedures: Final report on diagnostic efficacy*. Chicago: Efficacy Study Committee of the American College of Radiology.

Murphy, A. H. (1981). Subjective quantification of uncertainty in weather forecasts in the United States. *Meteorol. Rdsch.*, 34, 65–77.

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

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A, 135, 370–384.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice–Hall.

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

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39, 806–820.

RECEIVED: October 25, 1985