APPLIED COGNITIVE PSYCHOLOGY, VOL. 8, 621–639 (1994)

A Model for Predicting the Outcomes of Basketball Games

EVAN HEIT*
Northwestern University, USA

PAUL C. PRICE
University of Michigan, USA

GORDON H. BOWER
Stanford University, USA

SUMMARY

We investigated the task of predicting the outcomes of sporting events, in particular, basketball games. In two experiments, college students predicted the outcomes of a series of National Basketball Association (NBA) games. Following each prediction, the subject received feedback in the form of the actual outcome of the NBA game. After a period of initial learning about the relative strengths of the teams, the subjects were surprisingly successful in their predictions. We examined expert–novice differences by comparing the published predictions of professional oddsmakers to the predictions of the experimental subjects. The average predictions by the subjects were approximately as accurate as the predictions of experts. Finally, we applied a mathematical model, developed originally as an account of simpler learning experiments, to the subjects' responses. We found that the course of the subjects' learning about the teams was well-described by this model.

People are able to learn about the structure of a domain by putting together sources of partial information. For example, it is possible to discover the hierarchical structure of an organization or large family by observing pairwise interactions between members of the organization or family. Likewise, people can infer a linear ordering or ranking of objects along a single dimension by merely observing the outcomes of comparisons or contests between pairs of objects or participants. One type of linear order arises in dominance hierarchies, where the members of a set compete against each other for some scarce resource, and the winner is determined probabilistically from his or her quality along some dimension such as speed, power, or ability. The clearest cases of this kind occur in sports in which individual players or teams compete repeatedly, with an explicit winner and margin of victory. This description would apply to rankings of professional tennis players, boxers, track runners, and

This research was supported by an NIMH Individual Postdoctoral Fellowship to Evan Heit, an NSF Graduate Fellowship to Paul C. Price, and NIMH grant MH-47575 to Gordon H. Bower. Part of this research was presented at the 1991 Meeting of the Psychonomic Society, in San Francisco.

*Address for correspondence: Evan Heit, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208, USA.

CCC 0888–4080/94/070621–19

Received 27 October 1993
Accepted 14 January 1994

race horses, and to teams in baseball, basketball, or football. But similar pairwise contests and rankings arise in animal dominance hierarchies, in marketing wars between businesses trying to capture greater market share with competing products, and in political power or social status hierarchies.

We have previously developed a mathematical model to describe how people learn linear orders of participants by observing the outcomes of pairwise contests (Bower and Heit, 1992). In the present study, this abstract model was applied to the concrete task of a gambler or bookmaker learning to bet on professional basketball games. How do gamblers learn the relative strengths of basketball teams so that they can confidently bet on the winners and point spreads[1] for many games every season? To study this question empirically, we examined how college students learned to estimate winning point spreads as they progressed through a season's worth of basketball games, noting the winning team and the winning margin or point spread for each game. The stimuli in these studies were derived from actual games played during a National Basketball Association (NBA) season.

This research had several purposes. First, it was an attempt to apply a simple mathematical model of learning to a simulation of learning about the strengths of basketball teams and predicting the outcomes of games. Typically, such models are applied to tightly controlled experimental situations, in which the stimuli are contrived by the researcher, rather than to learning about part of the real world such as NBA teams. Thus, this research was intended to assess the generality of the model as a description of a real task. Second, the use of actual NBA games allowed us to evaluate the model, and our subjects, in terms of accuracy in predicting the true outcomes of the games. Can a group of college students learn to accurately make predictions of the winners of NBA games? How does the accuracy of the students compare to the accuracy of a mathematical model? Finally, we wanted to compare the subjects' judgements and the model's predictions to the forecasts of professional experts in published reports. Here we investigated whether the informational advantage enjoyed by the experts led them to be more successful than our relatively unknowledgeable subjects and our simple mathematical model.

We consider these studies to belong to the field of everyday memory research, because the task of predicting the winners of sporting events is a popular activity outside of the laboratory, and our stimuli were derived from real-world events. The learning in our studies was enacted by subjects in the laboratory, but that is the case in much everyday memory research (Gathercole and Collins, 1992). However, for comparison we also examined learning outside of the laboratory, by evaluating expert predictions reported in a syndicated newspaper column. These expert predictions are taken quite seriously by sports fans, and it would be quite surprising if our subjects in the laboratory were as accurate as the experts.

Considerations in developing the model

The model we propose comes out of the linear operator tradition of mathematical learning theory (Bush and Mosteller, 1955; Estes, 1972), which would typically predict only participants' win probabilities. In addition, our model is intended to predict

[1] The *point spread* of a game is defined as the difference between the number of points scored by the winner and the number of points scored by the loser.

people's estimates of the number of points by which any team in a league would win or lose to any other team in the league.

What properties of dynamic judgements should the mathematical model try to capture? We propose that when people learn orderings from specific outcomes, they bring to bear simplifying assumptions in order to reduce the complexity of the task. First, we suggest that people assume that they are indeed learning about objects or participants that are well-described by an ordering along a single dimension, such as strength or power. Some evidence for this proposal is provided by research on the learning of social structures from sets of propositions such as 'person A influences person B', 'person C influences person A', and so on (for a review, see Smith and Mynatt, 1982). These studies have shown that it is easier for people to remember a set of propositions if they permit a linear ordering of the elements than if they do not.

Furthermore, it is assumed that there is a monotone relationship between the strength of a participant and its ability to compete in contests. In particular, the likelihood of participant A winning a contest over participant B would be some monotone increasing function of the difference (or ratio) of their strengths. These assumptions imply that triplets of comparisons must follow stochastic transitivity, that is, if P(A wins over B) > 0.5, and P(B wins over C) $\geqslant 0.5$, then P(A wins over C) > 0.5. This transitive property of orderings implies that strength values are context-independent; that is, the strength of a participant is the same regardless of the other participants in the contest. Thus we can learn about the relative orderings of participants A versus C by noting how each fares against common opponents. So if A usually wins over B, and B usually wins over C, then the prediction that A will win over C is warranted. By assuming transitivity, the learners' task is greatly simplified: They need only to remember n strength values to rank n participants, rather than $n(n-1)/2$ context-dependent strength values that would be needed to predict outcomes if each participant had different strengths depending on who else was in the contest.

In addition, we wanted the model to have a recency basis. That is, in updating the strength of a participant, recent outcomes should have a greater influence than remote outcomes. This recency bias is consistent with numerous results in memory research showing forgetting over time. Recency bias is also consistent with moving average models of serial judgments, such as repeated predictions of the current price of a stock on the basis of its past values (Andreassen, 1991). Remembering recent events better than remote events could be an adaptive strategy for predicting outcomes, because more recent events may be more relevant to present needs (Anderson, 1990).

Finally, and most importantly, the model's estimate of a team's strength tends to rise when it wins and fall when it loses. However, not all wins or all losses convey the same information. Specifically, beating a strong team implies having more strength than it has, and losing to a weak team implies less strength than it has. The need to take into account such comparative information was demonstrated by Neely (1982). Neely's study used electoral races between political candidates rather than athletic teams. During a training phase, subjects learned that candidate A beat candidate B in 80 per cent of their races; when candidate X ran against candidate A, they each won 50 per cent of those races; and when candidate Y ran against B, each won 50 per cent of those races. The crucial question posed to subjects in a

later test phase was how well candidate X would fare against Y. If subjects evaluate candidates' strengths only by their percentage of past wins and losses, then X and Y should be equivalent. However, if they also attend to the strength of the opponents that a candidate beats, then a different outcome is expected: X should be judged as stronger than Y. In fact, Neely's subjects showed such a preference, choosing X over Y by a two to one margin in a contest between X and Y. Interestingly, analogous results to those of Neely (1982) have been obtained with animal subjects. Zentall and Sherburne (1994) trained pigeons on two simultaneous discriminations between coloured keys. When presented with stimuli A and B, the pigeons were always reinforced for pecking A, but never reinforced for pecking B. When presented with stimuli C and D, the pigeons were reinforced half the time for pecking C, but never reinforced for pecking D. When the pigeons were subsequently presented with stimuli B and D, neither of which had ever been reinforced, they showed a strong preference for B, the stimulus that had been paired with the stronger opponent (see also Belke, 1992; von Fersen, Wynne, Delius, and Staddon, 1991).

The adaptive comparison model

We now present a mathematical model that incorporates these considerations (Bower and Heit, 1992). We have previously applied this model to simpler tasks such as classic two-armed bandit studies from learning research (e.g. Atkinson, 1962). Although this model is applicable in a variety of real-world competitive situations, in this paper we apply it only to basketball games. This domain is especially appropriate because basketball teams play repeated contests against each other, and because each outcome is easily summarized in terms of the winner, loser, and point spread of the game.

The model assumes that each team, i, on trial n, is described by a value, $s_i(n)$, on a strength scale, with better teams having higher strengths. Trial n refers to the nth game played in a season. We assume that, in a game between team i and team j on trial n, the point spread may be predicted from the difference, $s_i(n) - s_j(n)$. If this difference is positive, then the difference corresponds to the number of points by which team i is predicted to defeat team j. If this difference is negative, then the difference corresponds to the number of points by which team j is predicted to defeat team i.

It is assumed that all teams start the season with strengths of zero. In the simplest version of the model, when team i plays team j in game n, team strengths are adjusted according to equations 1 and 2:

$$s_i(n+1) = s_i(n) + \theta[\text{outcome}(n) - \{s_i(n) - s_j(n)\}] \quad (1)$$

$$s_j(n+1) = s_j(n) + \theta[-\text{outcome}(n) - \{s_j(n) - s_i(n)\}] \quad (2)$$

Note that these equations are identical except for the sign of the *outcome*, which has the absolute value of the actual point spread. This *outcome* will be given a positive sign if team i wins and a negative sign if team j wins. (Assignment of particular teams to the letters i and j is arbitrary, and does not affect the use of the equations.) Finally, other teams, k, that do not play in game n do not change their strengths, that is, $s_k(n+1) = s_k(n)$.

This model has one free parameter, θ, which represents the learning rate. If $\theta = 0$, then no learning will take place and subjects' estimates of team strengths will

never change. If $0 < \theta \leqslant 1$, then the strengths will be learned, so that a team's strength will increase if the outcome is higher than expected, and will decrease if the outcome is lower than expected. In the case that two teams play each other repeatedly, the asymptomatic expected value of $(s_i - s_j)$ will equal the average outcome of all games played between the two teams (Bower and Heit, 1992).

This model also embodies the assumption that recent outcomes have a greater influence on a team's current strength than do remote outcomes. One can see why by rearranging equation 1 to produce:

$$s_i(n+1) = \theta(\text{outcome}(n)) + \theta(1-\theta)(\text{outcome}(n-1)) + \theta(1-\theta)^2(\text{outcome}(n-2)) + \ldots$$

(This equation would apply to the simple case in which team i repeatedly plays teams of zero strength.) Thus, so long as $0 < \theta < 1$, outcomes from the remote past are gradually forgotten over time (cf. Busemeyer, 1991).

This model also embodies the assumption that the adjustment of team i's strength depends on the strength of its opponent. A team gains strength when it wins by a larger margin than expected, and loses strength when it is defeated by a larger margin than expected. Because the expected margin of victory or defeat is a function of the absolute strengths of both teams involved, both determine the change. This independence implies that a team can gain strength, even in defeat, when playing against strong teams, and can lose strength, even in victory, when playing against weak teams. This implication is consistent with Neely's (1982) findings reviewed earlier.

The actual equations used to describe the present experiments were more general versions of equations 1 and 2, shown as equations 3 and 4. This version of the model will be referred to as the adaptive comparison model:

$$s_i(n+1) = s_i(n) + e^{-\kappa n}\theta[\text{outcome}(n) - \{s_i(n) - s_j(n)\}] \quad (3)$$

$$s_j(n+1) = s_j(n) + e^{-\kappa n}\theta[-\text{outcome}(n) - \{s_j(n) - s_i(n)\}] \quad (4)$$

The only change is the addition of the free parameter κ and the $e^{-\kappa n}$ term, which allows the net learning rate to decrease as the trial number increases, if $\kappa > 0$. Note that if $\kappa = 0$, then these equations are the same as the previous ones. The motive behind this addition is that learning might become slower, or changes in strength might become more conservative, as the learner has observed more outcomes. A similar assumption about decreasing learning rates is common for connectionist network models (Hinton and Sejnowski, 1986).

EXPERIMENT 1

In this experiment, we evaluated the adaptive comparison model as an account of people's learning about strengths of basketball teams, by simulating the course of a season's games between six professional teams. To ensure that the subjects responded on the basis of the outcomes of these games, rather than from prior beliefs about specific teams, we disguised the names of the teams. Otherwise, the game schedule, including the outcomes, was derived from the 78 games played between the six teams in the Atlantic Division of the NBA during the 1989–90 season. For each game, a subject predicted the winning team as well as the point-spread difference

between the winner and the loser. After the prediction was recorded, the actual outcome was presented in terms of the names of the winning and losing teams and the point spread. Thus, we simulated the real-word situation of predicting game outcomes and then learning of the actual outcomes from media sources, rather than simulating the entire experience of watching 78 basketball games. The final score (that is, the total number of points accumulated by each team), was not presented, but that information is usually less important to gamblers and fans than the winner and the point spread. In addition, other information, such as which team in a contest was playing on its home court, was not provided in this experiment.

Method

Subjects

The subjects were 25 University of Michigan undergraduates (12 female and 13 male) who participated in partial fulfillment of an introductory psychology course requirement. The subjects completed a pretest questionnaire to assess their knowledge and experience with basketball, such as how many games they had attended or played during the past year and whether they considered themselves basketball fans.

Stimuli

The background stimuli were fictional descriptions of six basketball teams. Each team was described by its name, the name of its coach, its colours, and the name of its home arena. Each description was formed by randomly combining one team name, coach name, set of colours, and arena name into a description of the form: 'the (team name) are coached by (coach name). Their colours are (team colours), and they play in (arena name)'. For example, the description of one team may have been 'the Piranhas are coached by Norm Jackson. Their colours are red and green, and they play in Morris Coliseum'. These materials are summarized in Table 1. The background test, administered after subjects had read the background stimuli, consisted of six multiple-choice questions, asking for one piece of information (i.e. coach, colours, or arena) about each team. The background stimuli and test were intended to make the task more realistic and interesting, and to improve subjects' memories for game outcomes by making the teams distinctive from each other.

Table 1. Descriptive information for teams

Team names	Coaches' names	Team colours	Arena names
Buffaloes	Norm Jackson	Blue and cream	Alexander County Stadium
Crocodiles	John 'Bubba' Levy	Brown and white	The Great Pavilion
Foxes	Mike Milhouse	Green and black	Lee Fieldhouse
Piranhas	O. P. Rodriguez	Red and green	Morris Coliseum
Ravens	Fred 'Happy' Wilson	Scarlet and yellow	The Queendome
Vipers	J. B. Winkles	Silver and purple	Webster–Stein Arena

During the training phase of the experiment, subjects predicted the outcomes of 78 basketball games. The outcomes of these games were those of games played between the six teams of the Atlantic Division of the NBA during the 1989–90

season, presented in the order in which they originally occurred. To avoid effects of subjects' prior knowledge about the real teams, the names of the NBA teams were replaced with the names of the six fictional teams presented as background stimuli. The only outcome information presented during the training phase was the winner of the game, the loser, and the number of points by which the winner was ahead (the point spread).

During the transfer phase, subjects predicted the outcomes of the 15 basketball games obtained from each possible pairing of teams; no outcome information was given. (The transfer phase might be described as a simulated set of end-of-season playoffs.) The transfer pairs were presented in a different random order for each subject.

Procedure

Subjects were run in groups of one to four. Stimulus presentation and response collection were controlled by IBM PC-compatible computers running a program written with Micro Experimental Laboratory software (Schneider, 1988). First, subjects were presented with the background stimuli, which they were instructed to study for a few minutes in preparation for a memory test. Immediately after reading the final team description, they completed the background test.

Before beginning the training phase of the experiment, subjects read instructions (reproduced in the Appendix). On each trial, the names of the competing teams were presented on the computer monitor. The names were centred on the display, one above the other, with the abbreviation 'vs' between them. The top or bottom position of the team names was determined randomly on each trial for each subject. The subject predicted the winner of the game by pressing either the up- or down-arrow key to highlight the predicted winner. The question, 'By how many points do you think the (predicted winner) will win?', then appeared on the display, and the subject typed a number from 1 to 99. Upon pressing the enter key, the display was cleared and the subject's prediction was again presented along with the actual outcome of the game, that is, the winning and losing teams and the point spread. The subject then pressed the space bar to initiate the next trial.

Immediately after the training phase ended, the transfer phase began. Subjects read further instructions (as reproduced in the Appendix), and then predicted the outcomes of the 15 transfer games.

Results

Evaluation of model

The model was evaluated in terms of its fit to the average responses of the subjects. The average response on a training trial was the mean point-spread estimate from all 25 subjects. A positive sign was assigned to a subject's estimate when the subject had correctly predicted the winner and a negative sign was assigned to the estimate when the subject's prediction of the winner was incorrect.

The predictions of the adaptive comparison model were obtained by treating the model as a simulated learner in the experiment. For each training trial, the model's prediction was determined from the current strength estimates for the two teams that played. After a prediction was made for the trial, these strength estimates were updated according to equations 3 and 4, taking into account the actual outcome

of that game. The model has two free parameters, θ and κ, representing the initial learning rate and the decrement in the learning rate per trial. A simulated series of games was run for a given pair of parameter values, and the model was evaluated in terms of the mean squared error of prediction. That is, the squared difference between the model's prediction and the subjects' average response was computed for each of the 78 trials, then these 78 squared differences were averaged. Successive combinations of values for these two parameters were evaluated iteratively, according to the minimization procedure of Chandler (1965). The final parameter estimates that were chosen minimized the mean squared error of prediction of the model. For $\theta = 0.084$ and $\kappa = 0.024$, the correlation between the model's predicted point spreads and the subjects' average responses was quite good, $r = 0.83$. The average absolute value of the error of prediction by the model, on each trial, was only 2.2 points. Figure 1 shows the average responses and the model's predictions over the course of the season. The model's predictions appear to track the subjects' responses quite well.

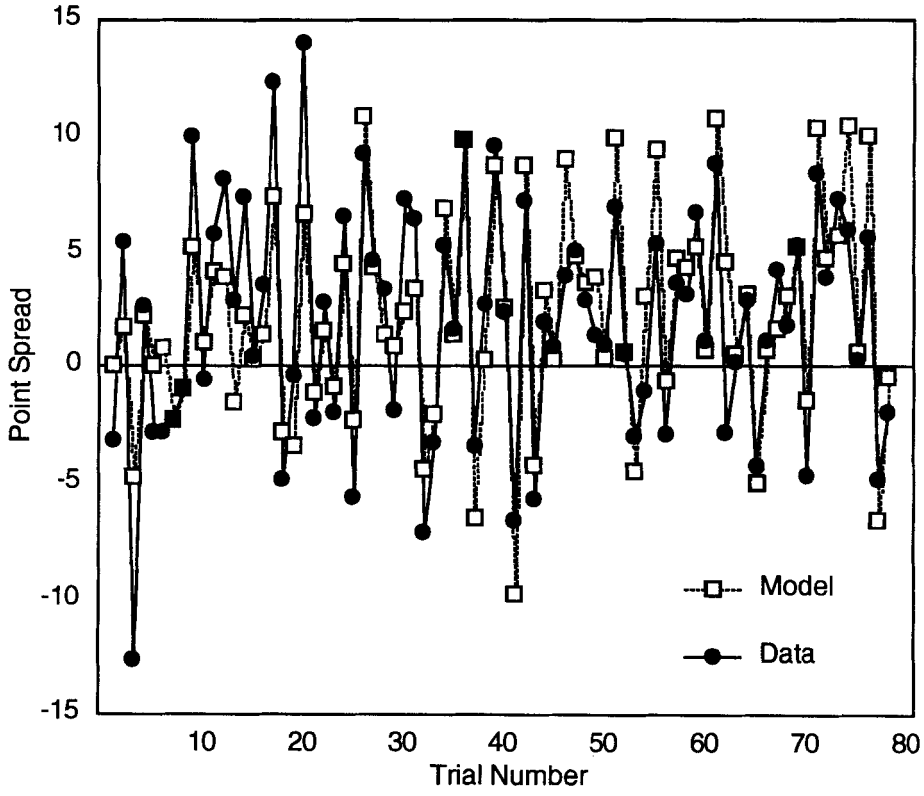

Figure 1. Adaptive comparison model predictions and average subject point spreads on each trial, experiment 1.

It was expected that the model would fit better to the latter part of the season, because earlier responses would be largely composed of guessing. Indeed, the fit

of the model was even more impressive for just the last 39 games. Using the same predictions obtained by fitting the model to the whole season, the correlation between the model's point spreads and the responses for the second half of the season was 0.89, and the average absolute error was only 1.8 points.

The learned strengths of the six teams, as obtained from the best-fitting version of the model after the simulated training trials, were used to predict additional point spreads for the 15 transfer games. (Because no feedback was provided during the transfer phase, it was assumed that no further learning took place.) Again, the adaptive comparison model fared well; the correlation between the model's predicted point spreads and the average responses was 0.82, and the average absolute error of the model was only 1.6 points. Just considering the winner rather than the point spreads, the model predicted the same winner as did the subjects' averages in all 15 transfer games.

Accuracy of predictions

Clearly, the subjects in this experiment responded systematically, and in a manner that was well-captured by the adaptive comparison model. But how successful were the subjects at predicting the actual outcomes? Note that these two issues are distinct; it would be possible for subjects to be systematic in their responses, but still not be very accurate. For a standard of comparison, we examined the accuracy of experts, in particular the published predictions of Tribune Media Services, which appear in the *Chicago Tribune* and are syndicated in many other newspapers. The predicted winners and point spreads for the games of the day appear in the morning newspaper; such predictions are often referred to as the 'morning line' or the 'Vegas line'. We presume that the professionals who made these predictions had access to much more information than we provided to our subjects, such as information about particular players and team histories. In addition, professional handicappers would have access to external memory, such as written records of game outcomes. In contrast, our subjects had to rely on their own memories. Furthermore, our subjects only observed the outcomes of the 78 intradivisional games. Each team also played approximately 55 games outside of the Atlantic Division. These outcomes might have also helped the experts to infer the strengths of the teams. Thus, the accuracy of the experts should be considered an upper limit or ideal to which our subjects and our model might aspire.

Here, we describe the accuracy of predictions for the second half of the season.[2] The subjects were, of course, less accurate during the first half, because they were initially learning about these teams and were guessing to a large extent. First, accuracy was evaluated in terms of the correlation between the actual point-spread outcomes of the games and the predicted point spreads (of the experts, the average subject, and the model). Surprisingly, the experts, our subjects, and the adaptive comparison model were about equally successful; the r values over the 38 games were 0.41, 0.47, and 0.42, respectively. Next, accuracy was assessed in terms of the average absolute error, in points. Again, the experts, the subjects, and the model were comparable, with average absolute errors of 9.4, 9.9, and 9.1 points per game, respectively. Third, we assessed accuracy just in terms of predicting the winning team, rather

[2] We evaluated accuracy for 38 of 39 last games of the season. One game was dropped because it was presented out of sequence during the experiment due to typographical error by the researchers.

than predicting the point spread. The adaptive comparison model was incorrect on eight games, and the experts and the subjects were incorrect ten times each.

Note that we assessed the accuracy of the subjects' predictions in terms of the group average; the predictions of individual subjects were generally less accurate than the group as a whole. Comparing individual subjects' predicted point spreads to the actual point spreads, we found that the r values for the second half of the season had a mean of 0.18 and ranged from -0.31 to 0.39. We assumed that an individual's judgements would be affected by random error that was eliminated by averaging across subjects.

Finally, whether or not the subjects considered themselves basketball fans had no discernible effect on the subjects' accuracy, nor on the model's fit to their data.

Discussion

We found that the adaptive comparison model gave a quite good account of how the subjects learned to predict the outcomes of NBA games. To our surprise, these experimental subjects were as successful at predicting the outcomes as were experts who made nationally-published predictions. Furthermore, these expert predictions were only modestly successful, with a correlation of about 0.4 with the actual point spreads, an average error of about 9 points per game, and a correct prediction of which team would win on about 70 per cent of the games. At least for this series of games, the potentially unlimited sources of additional information available to the experts did not benefit their predictions. Of course, a given series of contests may contain upsets and other unexpected outcomes. Therefore, we attempted to replicate these results in another experiment.

Out of fairness to the experts, we note that there are different possible success criteria for professional predictions of point spreads. Certainly, one criterion with much face validity is that the professional predictions should resemble the actual outcomes. However, professional point spreads may serve other functions. For example, a bookmaker may try to set a point spread to generate equal amounts of money bet on the two competitors, that is, a 'balanced book' (Abt, Smith, and Christiansen, 1985). The balanced-book criterion for setting a point spread is not exactly the same as the criterion of predicting the outcome that is considered most likely, but they are closely related. The creation of a balanced book suggests that a large number of knowledgeable sports bettors, taken as a group, find the line to be a reasonably good prediction of the game outcome.

EXPERIMENT 2

This study was a replication of the previous experiment with a schedule featuring different teams. In experiment 1, the success of the subjects, and the success of the adaptive comparison model, might have been due partly to particular details about the games selected. In addition, the adaptive comparison model was extended for experiment 2 to take account of an additional piece of valid information in making its predictions and updating team strengths. In the domain of basketball, gamblers are likely to use their knowledge of relative strengths of teams as well as the fact that teams generally play better when they are on their home courts.

We assumed that our subjects would integrate home-court information with their knowledge of the teams' strengths by adding a constant value to the perceived strength of the home team. Thus, the expected outcome of a game between home team i and visiting team j is $s_i(n) - s_j(n) + h$, where h is a positive constant reflecting the home-court advantage. In addition, the learning equations for the adaptive comparison model were modified to incorporate the home-court advantage parameter, as shown in equations 5 and 6:

$$s_i(n+1) = s_i(n) + e^{-\kappa n}\,\theta\,[\text{outcome}(n) - \{s_i(n) - s_j(n) + h\}] \tag{5}$$

$$s_j(n+1) = s_j(n) + e^{-\kappa n}\,\theta\,[-\text{outcome}(n) - \{s_j(n) - s_i(n) - h\}] \tag{6}$$

Method

Subjects

The subjects were 25 undergraduates at the University of Michigan (14 female and 11 male) who participated in partial fulfillment of an introductory psychology course requirement.

Stimuli

The stimuli were the same as in experiment 1, except for the following changes. The outcomes to be predicted during the training phase of the experiment were those of the 70 games played among six of the seven teams in the Midwest Division of the NBA during the 1989–90 season. Also, to allow the use of biasing information, the home team was designated by an 'H', and the visiting team by a 'V', displayed next to each team's name.

After training, 45 transfer trials were given. During the first 30 transfer trials, each one of the 15 possible pairings appeared twice, once with each team designated as the home team. These 30 pairings appeared in a different random order for each subject. For the final 15 transfer trials, each possible pairing of teams appeared once, but no home-team information was provided. These questions were intended to assess how much knowledge the subjects had of team strengths independent of home-court bias. These 15 trials appeared in a different order for each subject.

Procedure

The instructions preceding the training phase of the experiment included the following paragraph:

> The home team will always be designated by an 'H' after its name, while the visitor will have a 'V' after its name. You may want to take into account which team is the home team when you make our predictions.

The procedure, in other respects, was the same as in the first experiment.

Results

Evaluation of model

The home-court version of the adaptive comparison model, in equations 5 and 6, was evaluated in terms of its fit to the average point spreads predicted by subjects. The values of the three free parameters were estimated so as to minimize mean-squared

error over the 70 training trials, using a procedure analogous to that of experiment 1. The parameters θ and κ were 0.14 and 0.21, respectively, and the home-team advantage parameter, h, was 3.97 points. The correlation between the model's predicted point spreads and the average responses of subjects was very good, with $r = 0.88$, and the average absolute error of prediction on each trial was just 1.7 points. For just the second half of the season (the last 35 games), the fit of the model was even more impressive. Using the same predictions obtained by fitting the model to the whole season, the correlation between the model and the data for the second half was 0.93, and the average absolute error was only 1.3 points. Figure 2 shows the close correspondence between average responses and the model's predictions over the course of the season.

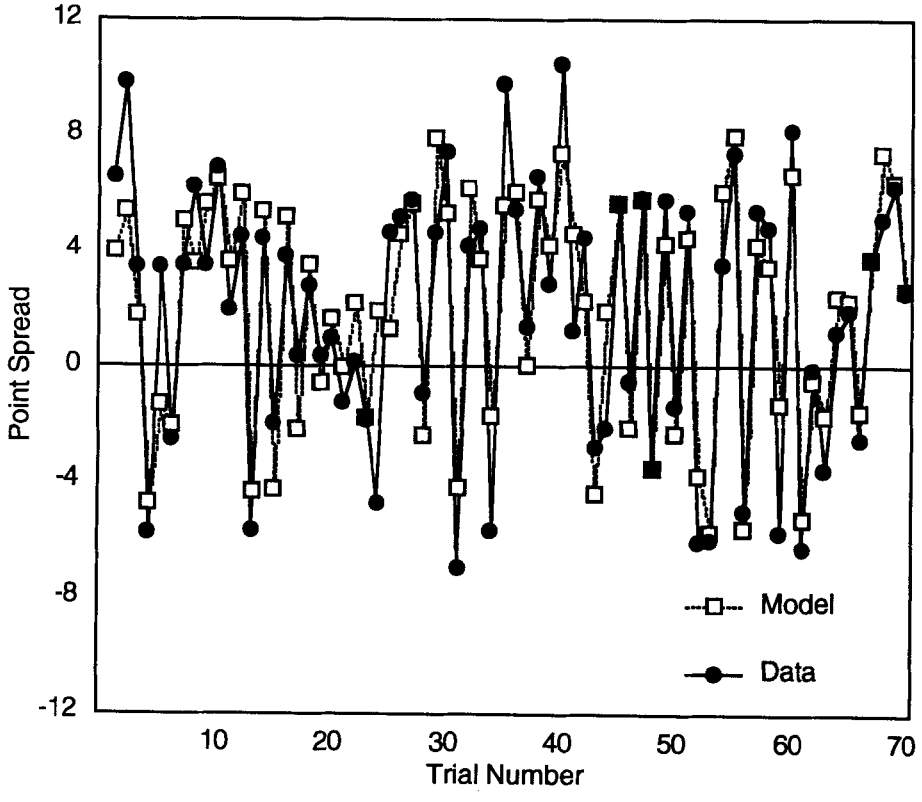

Figure 2. Adaptive comparison model predictions and average subject point spreads on each trial, experiment 2.

The learned strengths of the six teams, obtained by fitting the model to the training trials, and the estimated home-team bias parameter, were used to make further predictions for the transfer games. Over the first set of 30 transfer trials, for which home-court information was provided, the fit of the model was excellent for point spreads, $r = 0.97$ and the average error was 1.0 points. The model predicted the same winner as did the subjects' average point spreads in 26 of the 30 trials. For the final set of 15 transfer trials, during which the home team was not indicated, the correlation

between the model and the data was 0.85 and the average error was 1.5 points. The model predicted the same winner as did the subjects' average point spreads in 12 of the 15 trials. The results on the final set of transfer trials indicate that subjects were responding systematically to perceived team strengths, and that subjects had not merely learned to make predictions on the basis of home-court advantage.

In an additional analysis, the model was fitted separately to average responses of subjects who considered themselves basketball fans and subjects who were not fans. The model fit about equally well to judgments made during the training phase by the two groups of subjects. Interestingly, the home-team bias parameter was found to be 4.52 points for the fans and 3.40 points for the non-fans. This result suggests that fans have a stronger belief in the impact of being the home team than non-fans.

Accuracy of predictions.
We assessed accuracy of predictions for 34 of the last 35 games of the season. (One game was dropped because the expert predictions were not published that day.) First, accuracy was evaluated in terms of the correlation between the predicted point spreads and the actual outcomes. As in experiment 1, the experts, our subjects, and the adaptive comparison model were about equally successful; the r values were 0.52, 0.49, and 0.49, respectively. Next, accuracy was assessed in terms of the average absolute error in points. Again, the experts, the subjects, and the model was comparable, with average absolute errors of 6.4, 7.1, and 6.7 points, respectively. Third, we assessed accuracy in terms of predicting the winning team. The experts were incorrect 7 times, the subjects were incorrect 12 times, and the model was incorrect 10 times. This particular result was the only measure over the two studies in which the subjects did appreciably worse than the experts.

Considering the responses of individual subjects rather than the average response of the group, we found that the r values for the second half of the season had a mean of 0.25 and ranged from -0.15 to 0.53. As in experiment 1, the average of the subjects' responses had a higher correlation with the actual point spreads (0.49) than did the responses of most of the individual subjects.

Discussion

As in experiment 1, the adaptive comparison model, now with an added bias parameter, provided a very good account of the subjects' predictions. Also as in experiment 1, compared to experts, both the subjects and the model were quite successful at the predicting the actual outcomes of these games. In general, predictions were somewhat more accurate in experiment 2. For example, the average error of prediction decreased from about 9 points to about 7 points. Part of this change may be attributable to differences between the two series of games; perhaps the truly weaker teams won more often in the series from experiment 1. However, it is clear that subjects in experiment 2 were taking advantage of home-court information, because the original adaptive comparison model, without the bias parameter, provided only a poor fit of their responses. For experiment 2, the model without home-court information had a correlation with subjects' average responses of only 0.24 on the training trials and 0.42 on the first set of transfer trials. (The average point errors for the model were 4.1 on training trials and 4.0 on transfer trials.)

GENERAL DISCUSSION

Additional studies

In two additional experiments, we carried out a direct evaluation of the theoretical assumptions underlying the adaptive comparison model. We constructed the stimuli for these additional experiments, rather than using NBA schedules. In one experiment, we pitted the adaptive comparison model against a simple averaging model, which also represented perceived strengths along a single dimension. However, the averaging model differs from the adaptive comparison model in that it uses all past outcomes equally for making predictions. In contrast, the adaptive comparison model is influenced more by recent outcomes. The critical manipulation in this study was that midway through the season, a very successful team began to lose most of its games, and a poor team began to win. On the transfer trials in this experiment, both the human subjects and the adaptive comparison model were especially influenced by the more recent performances of the two critical teams that had changed in success rate midway through the basketball season. In contrast, the averaging model gave a poor account of peoples' judgements on the transfer trials. Thus, the assumption of a recency bias is critical to the account of people's predictions.

In another experiment, we replicated Neely's (1982) result that people's inferences about strength are influenced not only by the success rate of a participant but also by the perceived strengths of its past opponents. To find direct evidence for this influence, we compared the adaptive comparison model to the absolute model (Bower and Heit, 1992). Learning in the absolute model is not influenced by the strength of opponents; otherwise the adaptive comparison and absolute models are identical. In the experiment, we set up the outcomes so that different pairs of teams had the same win–loss records and the same average margins of victory. But one team in each pair played against successful teams and the other team played against weak opponents. On the transfer trials in this experiment, both the human subjects and the adaptive comparison model predicted more success for teams that had played against strong opponents than for teams that had played against weak opponents. In contrast, the absolute model incorrectly predicted no such influence of strength of opposition.

Limitations of the adaptive comparison model

Although the adaptive comparison model has provided successful accounts of these experiments (as well as other experiments described by Bower and Heit, 1992), we suspect that it may have shortcomings. First, its assumption that options are represented along a unidimensional strength scale could be an underestimate of what people remember. Besides this unidimensional scale, people might also represent idiosyncratic information about particular pairings of teams. For example, team A might generally win its games and team B lose its games, but team B might consistently beat team A. (This phenomenon is often reported by sports fans, who claim that a poor team 'has the number' of a better team.) This example violates transitivity, because A is better than most teams, and most teams are better than B, but B is better than A. We have run pilot experiments using a design like this example, and we have found that people have considerable difficulty in learning to violate the unidimensional strength assumption; they generally continue to predict that A

will beat B. However, our preliminary results do not rule out the possibility that with extensive training, people would be able to learn that A is generally good, and B is generally bad, but B beats A.

A second potential limitation of the model pertains to the assumption that the strength of a particular team changes only after competitions involving that team. This assumption may be too strict. To see why, imagine that early in the season team A consistently defeats team B by small margins. Later in the season, team C consistently defeats team B by extremely large margins. The results of these later games involving B and C could conceivably lead subjects to adjust the strength of team A downward. That is, people might conclude that team A was not as good as it first appeared to be because the team it consistently defeated early in the season, team B, was even worse than it first appeared to be. That people do engage in such retrospective evaluation has been demonstrated by researchers applying simple associative learning models, much like the adaptive comparison model, to human contingency learning and judgment (Chapman, 1991; Shanks and Dickinson, 1987). Also, there exists some evidence that animals engage in retrospective processing (Matzel, Schachtman, and Miller, 1985).

A third anticipated limitation of the model is that in some situations, it might show much worse forgetting than people. Imagine that in the first phase of an experiment, team X wins consistently over team Y. Then in a second phase, only games between X and team Z are presented, in which Z consistently beats X. We expect that subjects would conclude that Z is stronger than X, and X is stronger than Y. However, with enough observations of Z beating X, the adaptive comparison model could predict that Y would beat X. The model would forget the remote outcomes in which X had consistently beaten Y, and use the more recent outcomes of X losing to assign a negative strength to X. Note that this problem would not occur if X continued to beat Y in the second phase of the experiment. This forgetting problem in our model resembles the catastrophic interference problem in connectionist network models (Ratcliff, 1990), in which the learning of new associations leads to an extreme amount of forgetting of old associations, unless the model is also presented with the old associations during the course of learning the new ones.

Extensions to the adaptive comparison model

So far, the adaptive comparison model has been evaluated in experiments using the domain of basketball games. For this domain, it seems appropriate to describe the difference in quality between two teams in terms of a difference in points, and to assume that there is a simple linear relation between strength differences and point spreads. However, further scaling assumptions would be required to apply the adaptive comparison model to other domains. For example, to learn about relative strengths of (American) football teams, the model might include the assumption that large point spreads are not much more informative than small point spreads. For example, say the football team B beats team A by 10 points, and that team C beats team A by 30 points. What is crucial about these outcomes is that both team B and team C won by substantial margins; it would not necessarily be appropriate to increase team C's perceived strength by three times as much as the increase in team B's perceived strength. Such knowledge about the point-spread scale might be incorporated into the adaptive comparison model by transforming the point

spreads logarithmically before the model is applied. Another domain for which additional scaling assumptions would be required is applying the model to outcomes of races, such as car races, horse races, or track events, in which the outcome would be described as the difference in finishing times between participants.

Applying the adaptive comparison model to races would require an additional extension. The model can predict the outcome of a race between multiple competitors, such as several horses or racing cars, using each of their estimated strengths, but the model is limited to learning only about outcomes between pairs of participants. People are surely not so limited; they can learn about relative strengths from outcomes of contests involving more than two competitors. One way to extend the model would be to treat a race between multiple competitors as a set of pairwise contests. For example, a finishing order of B D C A in a four-horse race would be treated equivalently to the pairwise outcomes (BD), (BC), (BA), (DC), (DA), and (CA). On a given trial, the learning equations of the model would be applied to each of these pairwise races to calculate the new strengths of the four horses. A problem with this method is that the model is non-commutative, so that the final strengths would depend on the order in which the hypothetical pairwise races were considered to occur. Therefore it is possible that a more elegant solution to this problem could be developed.

Using the adaptive comparison model to predict real-word outcomes

Our data allow us to test for the presence of bootstrapping, the phenomenon whereby a mathematical model (typically a linear model) of judges more accurately predicts outcomes than do the judges themselves (Camerer, 1981; Dawes, Faust, and Meehl, 1989). Our results suggest that the predictions of the adaptive comparison model were approximately as accurate as the subjects' means. However, we found that individual subjects were not as accurate as the group means or as the model. We would account for this finding by appealing to the traditional explanation of the bootstrapping phenomenon—namely, that individual judges are not consistent (Dawes, 1971). Subjects are influenced by a number of extraneous factors (e.g. variable effort, fatigue, and boredom) that introduce random noise into their predictions, thus lowering their overall judgmental accuracy.

It would not be necessary, however, to assemble a group of 25 subjects in order to use the model to make fairly accurate predictions about basketball games. Note that the adaptive comparison model (that is, equations 3 to 6) learns from the actual game outcomes, rather than from the subjects' responses. In fitting the adaptive comparison model to the subjects' responses, the subjects' responses were only considered in determining the free parameters of the model such as the learning rate. Alternatively, it is possible to choose these parameters arbitrarily, or better still, set them according to some evaluation of what has happened in past seasons. Thus, the adaptive comparison model can make predictions without any data from subjects.

This discussion leads naturally to the question of whether the adaptive comparison model might be used to improve on real-world predictions. For instance, could a savvy gambler use the adaptive comparison model to make more money when betting on the outcomes of basketball games? We found that the judgements of sports experts were no more accurate than our model, despite the many sources of information available to the professional oddsmakers. Indeed, such an informational edge pro-

vided to human judges does not always translate into superior performance (Dawes, 1971). Under some circumstances, the more knowledgeable the judges, and the more information they take into account in making their predictions, the poorer those predictions become. Yates, McDaniel, and Brown (1991), for example, studied predictions in a different kind of gambling context, that of predicting stock price changes (see also Stael von Holstein, 1972). They found that Masters- and PhD-level business students generally made poorer predictions than did business school undergraduates. Yates *et al.* (1991) attributed this result to the experts using various cues they incorrectly believed to be related to stock price changes, while the undergraduates relied on relatively simple judgement strategies. Although the adaptive comparison model was not found to be superior to experts, its performance is in some ways more impressive because it relies on such a limited range of information.

We developed our model strictly from psychological considerations for describing people's abilities to learn relative strengths and to make predictions. We were surprised to find that similar systems for making predictions have been introduced as normative models for sports fans. For example, the Stat-Key method for predicting the outcomes of sporting events (Sports Associates, Inc., 1991) recommends that fans predict the outcome, that is, the winner and point spread, of a game between two teams from the difference in their 'power ratings', which correspond to the s values in the adaptive comparison model. This estimate is then adjusted by adding or subtracting a number of points for the home-court advantage, as we did in applying the model in experiment 2. After the game is played, the Stat-Key method recommends that the power ratings be adjusted according to the difference 'between the estimated and actual differences in scores'. The power rating of the team that performed better than expected is increased, and the power rating of the team that did worse than expected is decreased. The adaptive comparison model updates team strengths in the same manner. Finally, according to the Stat-Key method, fans should decrease the degree of updating as the season progresses, just as the learning rate decreases over time in the adaptive comparison model.

To conclude, in our investigation of the task of predicting the outcomes of real professional basketball games, we found that our subjects' responses are well-described by a mathematical model that was originally proposed to explain learning in simple laboratory tasks. Interestingly, this descriptive psychological model is very similar to a normative model that has been prescribed for sports fans who wish to make accurate predictions (Sports Associates, Inc., 1991). Finally, in our evaluation of experts' published predictions for these same NBA games, we were surprised to find that the predictions of professionals were not appreciably more accurate than the predictions of the psychological model or of the average predictions of our subjects. This finding suggests that the relatively small amount of information provided to our subjects did not impair their ability to make real-world predictions, and correspondingly, the potentially unlimited sources of information available to experts did not make them any more successful than our subjects in the laboratory.

REFERENCES

Abt, V., Smith, J. F., and Christiansen, E. M. (1985). *The business of risk: Commercial gambling in mainstream America.* Lawrence, KS: University Press of Kansas.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Andreassen, P. B. (1991). Judgmental extrapolation and market overreaction: On the use and disuse of news. *Journal of Behavioral Decision Making*, **3**, 153–174.

Atkinson, R. C. (1962). Choice behavior and monetary payoffs. In J. Criswell, H. Solomon, and P. Suppes (Eds), *Mathematical methods in small group processes*. Stanford, CA: Stanford University Press.

Belke, T. W. (1992). Stimulus preference and the transitivity of preference. *Animal Learning & Behavior*, **20**, 401–406.

Bower, G., and Heit, E. (1992). Choosing between uncertain options: A reprise to the Estes scanning model. In A. F. Healy, S. M. Kosslyn, and R. M. Shiffrin (Eds), *From learning theory to connectionist theory: Essays in honor of William K. Estes*, Vol. 1 (pp. 21–43). Hillsdale, NJ: Lawrence Erlbaum.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory, Vol. 1: Cognition* (pp. 187–215). Hillsdale, NJ: Lawrence Erlbaum.

Bush, R. R., and Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, **27**, 411–422.

Chandler, J. P. (1965). STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science*, **14**, 81–82.

Chapman, G. B. (1991). Trial order affects cue interaction incontingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 837–854.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, **26**, 180–188.

Dawes, R. M., Faust, D., and Meehl, P. (1989). Clinical vs. actuarial judgment. *Science*, **243**, 1668–1674.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, **67**, 81–102.

Gathercole, S. E., and Collins, A. F. (1992). Everyday memory research and its applications *Applied Cognitive Psychology*, **6**, 461–465.

Hinton, G. E., and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1 (pp. 282–317). Cambridge, MA: MIT Press.

Matzel, L. D., Schachtman, T. R., and Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning & Motivation*, **16**, 398–412.

Neely, J. H. The role of expectancy in probability learning. (1982). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **8**, 599–607.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, **97**, 285–308.

Schneider, W. (1988). Micro Experimental Laboratory: An integrated system for IBM PC compatibles. *Behavioral Research Methods, Instruments, & Computers*, **20**, 206–217.

Shanks, D. R. and Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 21 (pp. 229–261). New York: Academic Press.

Smith, K. H. and Mynatt, B. T. (1982). Construction and representation of orderings in memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 26 (pp. 111–151). New York: Academic Press.

Sports Associates, Inc. (1991). *GamePlan College & Pro BASKETBALL Yearbook*, Vol. 14. Syracuse, NY: Sports Association Inc.

Stael von Holstein, C.-A. S. (1972). Probabilistic forecasting: An expriment related to the stock market. *Organizational Behavior and Human Performance*, **8**, 139–158.

von Fersen, L., Wynne, C. D. L., Delius, J. D., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, **17**, 334–341.

Yates, J. F., McDaniel, L., and Brown, E. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes*, **49**, 60–79.

Zentall, T. R., & Sherburne, L. M. (1994). Transfer of value from S+ to S− in a simultaneous discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, **20**, 176–183.

APPENDIX

Instructions preceding the training phase

In a recent basketball season, the six teams you just read about played a total of 78 games among themselves. Your task will be to predict the outcome of each of these 78 games. The games will be presented to you one at a time, in the actual order in which they were played, although you won't be told which team was the home team.

For each game, you will be asked to pick the winner and to guess the point spread. (The point spread is the number of points the winner will win by. For example, if team X beat team Y by a score of 110 to 100, the point spread was 10 points). It is important that you try to predict both the winners and the point spreads as accurately as possible.

Immediately after predicting the outcome of a game, you will be presented with the actual results of that game. This information should help you learn about the relative strengths of the six teams. Consequently, your predictions should improve as the season progresses.

Although you have as much time as you need to finish this experiment, don't get bogged down trying to predict the outcome of any one game. Just make your best guess, and move on. Also, please don't use paper and pencil to keep track of the games. Good luck!

Instructions preceding the transfer phase

For the final part of this experiment, you will be asked to predict the outcomes of 15 more games. Because these final 15 games are hypothetical—they have not actually been played—you will not be given any outcome information for them. Just predict the winners and the point spreads as best you can based on what you now know.