An Adaptive Strength Comparison Model of Predicting Basketball Winners

Evan Heit, University of Michigan

Paul Price, University of Michigan

and

Gordon Bower, Stanford University

Speech by G. Bower given at Psychonomic Society convention 1991

Evan Heit, Paul Price, and I have been interested in how people learn linear orderings of objects along a single dimension by merely observing the outcomes of comparisons or contests only between pairs of objects or participants. An example would be learning the ordering of heights of a classroom of children where one gets only pairwise comparative information. Our knowledge about rankings is frequently derived in just this manner, where we get pairwise comparative information but have to somehow integrate it all into an overall ranking of the objects.

One type of linear order arises in dominance hierarchies where the members of a set compete against one another for some resource, and the winner is determined probabilistically by his or her quality along some dimension such as speed, or power, or ability. The clearest cases of this kind occur in sports in which either individual players or teams compete against one another repeatedly, with a clear victor and a winning margin. This is the case for rankings of professional tennis players, boxers, track runners, stockcar racers, and race horses, and for 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 their competing products, and in political power, or social status hierarchies.

We have developed a mathematical model to describe how people learn such linear orders from observing the outcomes of pairwise contests. In order to situate this abstract model in concrete terms, we will apply it to the case of a gambler or a bookmaker who is learning how to bet on professional basketball games. How do bookies learn the relative strengths of basketball teams so that they can confidently bet on the winners and the point spreads for lots of games every weekend? Local newspapers carry listings of these point spreads for the upcoming games of the week; they're obviously set with considerable care and calculation, since the bookies' livelihood depends on it. Although we didn't have access to a group of naive bookies, we did the next best thing and solicited the aid of college students to serve as subjects. Thus, we examined whether naive college students could learn to estimate winning point spreads accurately as they progressed through a season's worth of games, noting the winning team and the winning margin or point spread for each game.

The model we developed to describe this learning comes out of the linear operator tradition of mathematical learning theory, which typically predicted only teams' win probabilities. Instead, our model will try to predict people's estimates of the number of points by which any team in a league will win or lose to any other team in that league. We think of this as a process of learning the relative strengths of the different teams. We surely want our model to have the property that its estimate of a team's relative strength rises when it wins and falls when it loses. However, not all wins and losses are equally informative. Specifically, if you beat a strong team, that tells us that you're probably stronger than they are; and if you lose to a weak team, that tells us that you're probably weaker than they are.

The need to take account of such comparative information was demonstrated in an experiment in 1982 by Jim Neely, the essentials of which are depicted in the first overhead (overhead 1). Neely used electoral races between political candidates rather than athletic teams but the principle is the same. During a training phase, college students learned that candidate A beat B in 80% of their races; also when candidate X raced against candidate A, they each won 50% of those races; and when candidate Y raced against B, each won 50% of those races. The crucial question posed to subjects in the test phase of the experiment was how well candidate X would fare against Y. Since X and Y had never run against one another, subjects have to draw an inference. If they evaluated candidates' strengths only by their percentage of past wins and losses, then X and Y should be equivalent. However, if subjects also take account of the strength of the opponents that a given candidate beats, then a different outcome is expected. On that basis, subjects know that candidate A is stronger than B, that candidate X is equally strong as A, and that Y is equally strong as B; therefore, they can infer that X is probably stronger than Y. And by and large that is the way that Neely's subjects chose by a 2 to 1 margin in the X versus Y contest.

We developed a model to explain these Neely results; the basic idea is that if entity A beats B, then the strength estimates for A and B are adjusted so that just that outcome would be predicted the next time they oppose one another. One neat test of that model is to see how well it does in accounting for people learning to predict a lot of contests such as a whole season's worth of games in the National Basketball Association. That idea led to the present experiments.

Let me describe our procedure. Twenty-five college students participated in a multi-trial prediction task (overhead 2). At the outset they were given the names and some concrete details about six fictitious professional basketball teams....the Ravens, the Rattlesnakes, the Foxes, and so on. An example is: "The Crocodiles are coached by Happy Wilson. Their uniforms are green and black, and they play in the Queendome (capacity 30,000)." After this, they then see a random sequence of 78 games in which the six teams play one another about 5 or 6 games. Each game is presented separately on the computer screen...such as "The Ravens play the Foxes"... and subjects type in who they think will win and by how many points, as shown in the bottom of the slide. The computer then gives feedback, such as "The Ravens win by 3 points," and 5 seconds later presents the next game to be predicted. At the end of 78 games with feedback, we ran a final test cycle through all 15 opponent pairs, obtaining a prediction but not providing any feedback about the outcome of those end games. For the training games we used the 1989-1990 season of actual game outcomes among the six teams in the NBA Atlantic Division, consisting of the Boston Celtics, New York Knicks, Washington Bullets, Philadelphia 76'ers, and so on. However, in order to keep our subjects naive, we disguised the real teams by fictitious names. We wanted to watch them learn throughout the course of the season, and see whether a simple math model would describe that process.

So, what is that math model? It is a linear operator learning model whose equations are shown in the third overhead. Here Si(n) and Sj(n) refer to the strength of teams I and J just before the nth game of the season. If they play against one another and I beats J by a point spread of size Pij(n), then we change the strength of I and J (the delta Si or Sj) so that their difference moves towards the observed point-spread, Pij(n). Here, theta is a learning rate constant between 0 and 1, on the order of .20 in our experiments; it gauges how rapidly the current strength difference is moved towards the outcome of the current game. If team I always beats J by 10 points, then this equation will adjust Si and Sj so that they eventually yield a difference of 10 points. This is shown by the equation at the bottom of the slide, showing how the difference changes towards the actual point spread, Pij(n). Incidentally, if team J beats I, then the same equations apply except now Pij(n) will be negative.

Several features of these equations should be noted. First, how much a win increases a team's strength varies directly with its strength of the opponent. Likewise, one's strength can be devastated by losing to a very weak team. This is analogous to the national collegiate football rankings where if a top-ten team is beaten by an unranked team, they are virtually dropped out of the top ten rankings. Second, the strength of a team can change only when it plays a game, although it may change its ranking because teams around it win or lose by a lot. Third, the change in strength directly reflects the point spread: thus, you gain more strength if you give a real shellacking to a strong team than if you barely eke out a win. Fourth, a team can win a game but nonetheless fall in the overall rankings if it beats a weak team but by far less than the expected point spread. All these are reasonable intuitions about the way sports fans and bookmakers adjust their beliefs about the relative strengths of teams or individual players in individual sports such as Pro tennis.

Over the course of a season, these equations adjust the strengths of the teams iteratively as shown in Overhead #4. Here just before game n the teams are ordered from A to F. In the nth game C beats A by 10 points, so those two strengths are adjusted. In the (n+1)st game E beats B by 5 points, so the strengths of E and B are adjusted. This diagram depicts a complete correction of the point spread which would arise if the learning rate, theta, were one; in fact, our subjects don't adjust their strength orderings nearly this much each time as this illustrates. However, the diagram does illustrate how the learning equations adjust the strength estimates iteratively as game after game of the season is encountered.

So, how well does this math model describe what our subjects do? Since every subject saw the same sequence of games, we can test the model by trying to fit the subjects' average prediction of the point spread for each game. For scoring purposes, we counted a subject's point-spread prediction as a positive number if he correctly predicted the winner of a game, and as a negative number if he chose the loser for that game.

Assuming equal initial strengths of zero, we estimated theta by a least-squares fit of the model to the average point spreads predicted by subjects. Those average point spreads are shown plotted over trials in Overhead #5. This graph shows that the model is rather faithfully following the bumps and dips of subjects' average predicted point spreads over the course of the season, including predicting the wrong team will win in many surprising games -- those are the games with scores below zero. The correlation between these two curves over the last half of the season is .90 with an average error of 2 points between the observed and model's values.

Another way to show the fit of the model is with the scatter plot (Overhead #6) of the model versus the subjects' average predicted point spreads over the last two games played between all 15 pairs of opponents. This shows a correlation of .96 which is pretty good for a simple model. A similar result holds for the final test trials which were given without feedback. The model predicts the same winner as do subjects in 14 out of 15 of those games.

Let's examine more closely the several ingredients of our model. The linear operator model can be viewed as taking a time-weighted average of a team's winning and losing margins, giving exponentially less weight to outcomes from games in the more distant past. One might ask whether that recency-weighting scheme is needed: perhaps we could do as well if we took the simple average, equally over all trials, of a team's winning and losing margins. The crucial difference between a simple averaging model and a recency-weighted model shows up if the strengths of some of the teams are reversed in mid-season. The simple averaging model would adjust only very slowly to the changed fortunes of the teams, whereas the recency-weighted model would adjust rapidly to their changed strengths.

Evan and Paul ran a 6-team experiment like this, where at midseason without warning to subjects the names of the best and worst teams were simply switched. As expected, the subjects and the recency-weighted learning model adjusted to this switch rather quickly, whereas the simple averaging model did very poorly in fitting the data following the midseason switch in strengths.

Let's return to the original experiment to examine another aspect of those data. Although the model does well in predicting the average point spreads that subjects estimate for each game, you may ask, how well are the subjects doing in predicting the *actual* point spreads? On a trial-by-trial basis, subjects are not doing very well nor is the model. For example, in Experiment 1, individual subjects predict the winner only 58% of the time. Although their average estimates predicted the winner 75% of the time, their estimates correlated only .21 with the actual point spreads. So just knowing the teams does not enable one to predict basketball game margins very well. The margins are just too variable.

One contributor to this variability is the home-court advantage: all sports fans know that teams play better when they play on their home court before their own fans. So we did a third experiment to see whether telling the subject which team was playing at home would enable them to improve their predictions. The experiment was similar to the first one with 25 subjects predicting the point spreads of 70 games, but before each prediction they were told the home team. To generalize our results, we used training games from a different league, namely, the Central Division of the NBA during the 1989-90 season.

To account for this home-court advantage, we altered the model in two ways as shown in Overhead #7: first, the predicted point spread was obtained after adding a constant H to the strength of the home team; second, the learning equations were altered, so that changes in strengths were calculated from the discrepancy of the outcome from the expected difference which included the home-court advantage.

The observed and predicted average point spreads over the season are shown in the overhead (#8). Both the model and subjects estimate the home-court advantage as worth about 4 points. Here the model is tracking the subjects' predictions fairly well, with an average error of 2 points. Over the last two meetings between the 15 pairs of teams, the model predicts the subjects' average point spread with a correlation of .93. Moreover, by knowing the home team, subjects become more accurate in predicting actual game outcomes, since the correlation of their average point spreads with the actual spreads increases from the earlier .21 up to .30.

However, a curious thing happens here: subjects' bias to bet on the hometeam is so strong that it almost blocks or overshadows the learning about specific team strengths. For example, the learning rate, theta, is less than a third here than what it was in Experiment 1; and the model does more poorly in predicting subjects point spreads during a final set of transfer trials when the teams all played on a neutral court where neither team enjoyed a homecourt advantage. That kind of overshadowing is interesting and we hope to investigate it further this kind of overshadowing.

We are currently investigating several extensions of the model and experiments. One extension (Overhead #9) examines people's inferences as they make comparisons across two different strength orderings. They first learn the strength ordering of two distinct sets of teams, ABC and DEF, then learn that the bottom team of one set (C) can beat the top team of the other set (D). The question is whether subjects can use that one bit of information to now completely rank the winners in contests between any pair of teams. A second extension (#2 in Overhead) will try to generalize the model to handle races between 3 or more contestants, say, between 3 horses, A, B, C, running a race in which the only outcomes reported are the distances between them at the finish line, say 4 and 2 lengths. One proposal is to alter the 3 strengths by treating this 3-way outcome as though it were a condensation of the outcomes of three pairs of races, as shown to the right of (2) in Overhead #9. This proposal assumes that the finishing margin between two horses is independent of what other horses are entered into the race...a kind of independence from irrelevant alternatives.

Thirdly, (Overhead #10), one defect of our current model is that it is deterministic, with no random elements in it. But of course, game outcomes are highly uncertain and probabilistic, which after all is why games are played. We could add day-to-day variability to each team's strength, yielding the picture shown in Item (3). This is like Thurstone's model of comparative judgment, which enables one to predict the probability of winning as well as the expected point spread in contests. A nearly equivalent approach is to convert strength differences into choice probabilities by taking the logistic transform, as shown.

A basic assumption of this model is that the contestants can be ordered along a single dimension of quality, goodness, or strength. But surely the world contains some cases in which the contestants are not consistently transitive. A simple illustration is shown in Overhead #9. Here teams ABC are equally good and teams DEF are equally bad, with any of the teams A, B, C usually beating all of the teams D, E, F. However, suppose that there is one weird exception whereby team E can consistently beat team C but none of the others. That state of affairs would violate the assumption of transitivity. We may ask whether subjects would be able to adjust to such a violation of transitivity, especially if it were a minor probabilistic violation embedded in a host of well-ordered, transitive probabilistic outcomes. In a pilot experiment with these arrangements, our subjects were just beginning to treat the C versus F game differently from all the rest by the end of a 150-game season. So with more extended training we would expect subjects to differentiate teams C and F depending on who they're playing. We are awaiting further data in that experiment.

For now, I think we can feel somewhat satisfied that our simple learning model has fared reasonably well in these first outings comparing it to experimental results.

- 1

Neely's 1982 Design

	80%	:	20%	
	A	>	B	
50/50%	‖		‖	50/50%
	‖		‖	
	X		Y	
TEST:	X	?	Y	
		68%	32%	

	Foxes	Rattlesnakes	Croc.	Buf.	Pir.
Ravens					
Foxes					
Rattlesnakes					
Crocodiles					
Buffalos					
Piranhas					

****

"The crocodiles are coached by Fred ("Happy") Wilson. Their uniforms are green and black, and they play in the Queendome (capacity 30,000)."

****

"The Ravens play the Foxes."

The winning team will be the _______.

They will win by _______ points.

If Team I beats Team J by $P_{ij}(n)$ points:

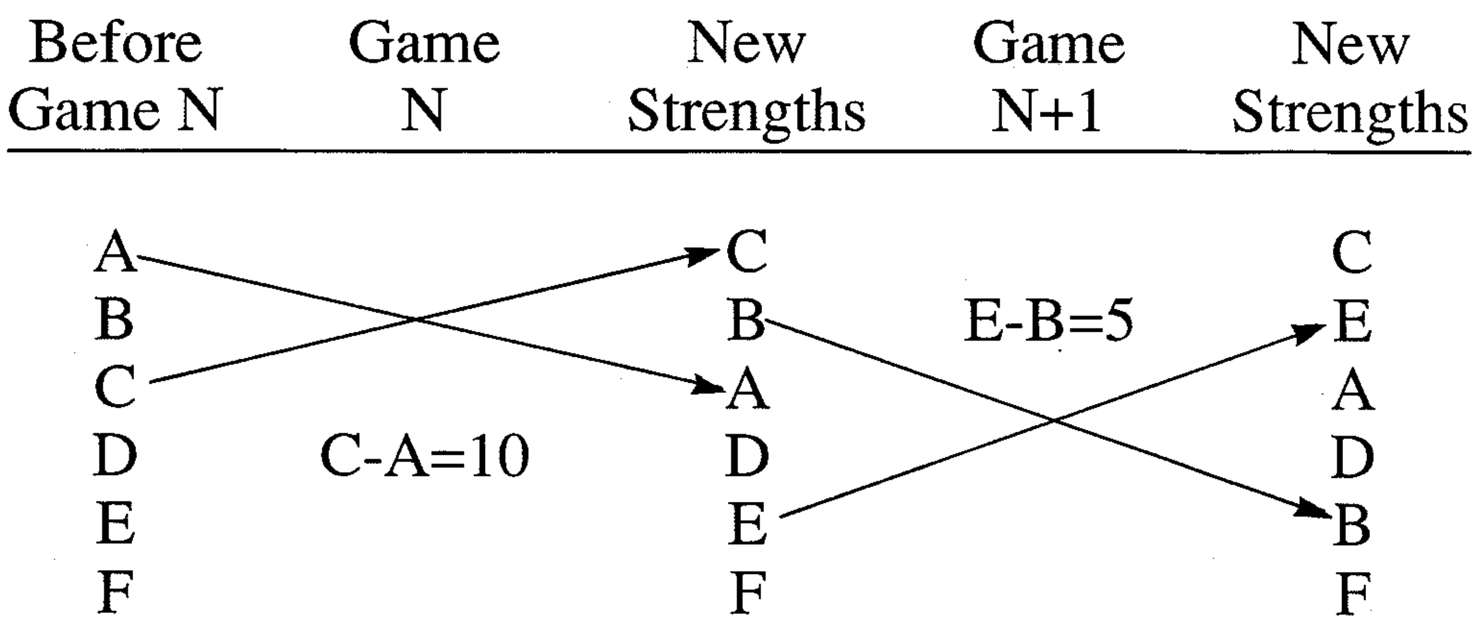
$$\Delta S_i = -\Delta S_j$$

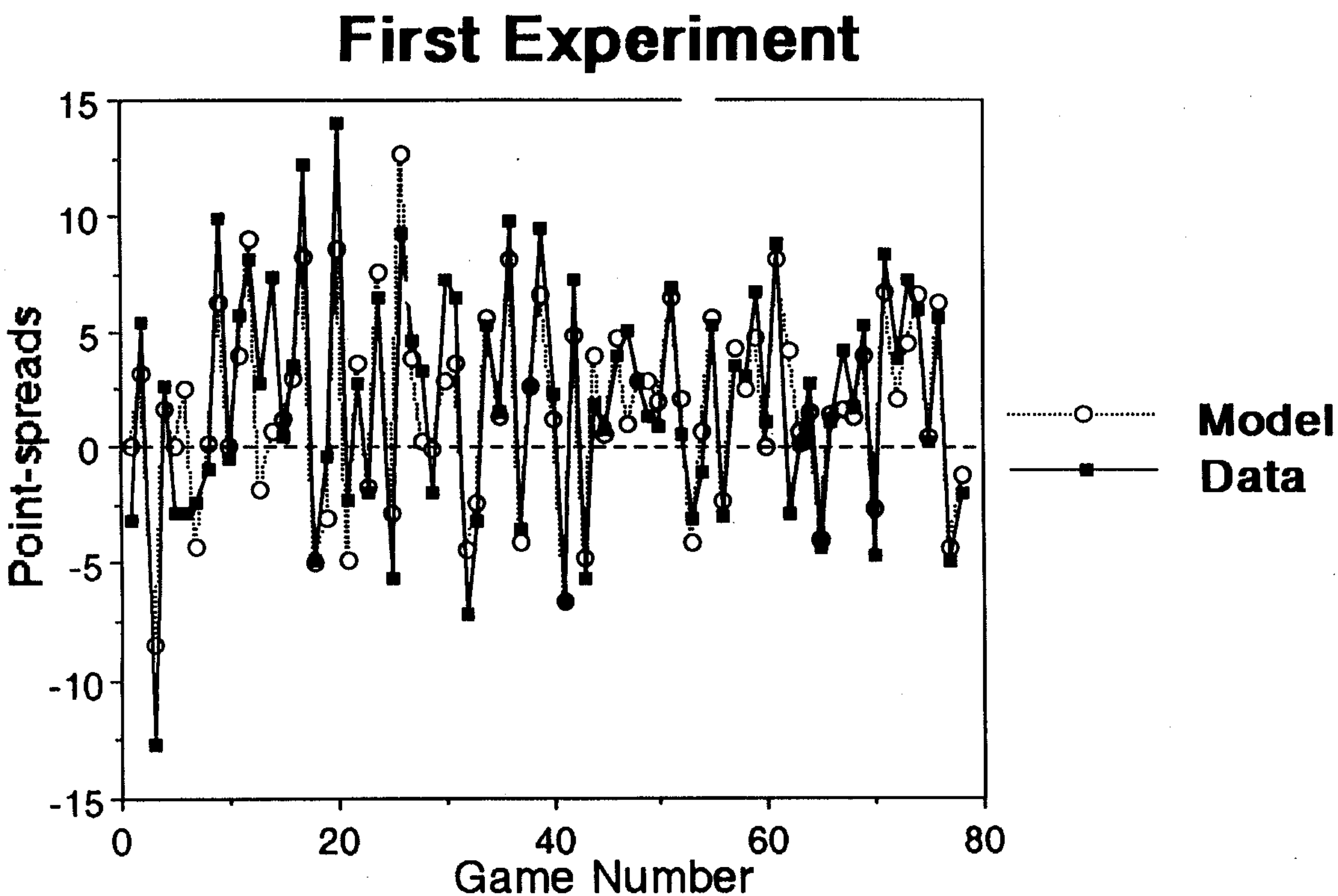
$$= .5\Theta\left[P_{ij}(n) - (S_i(n) - S_j(n))\right]$$

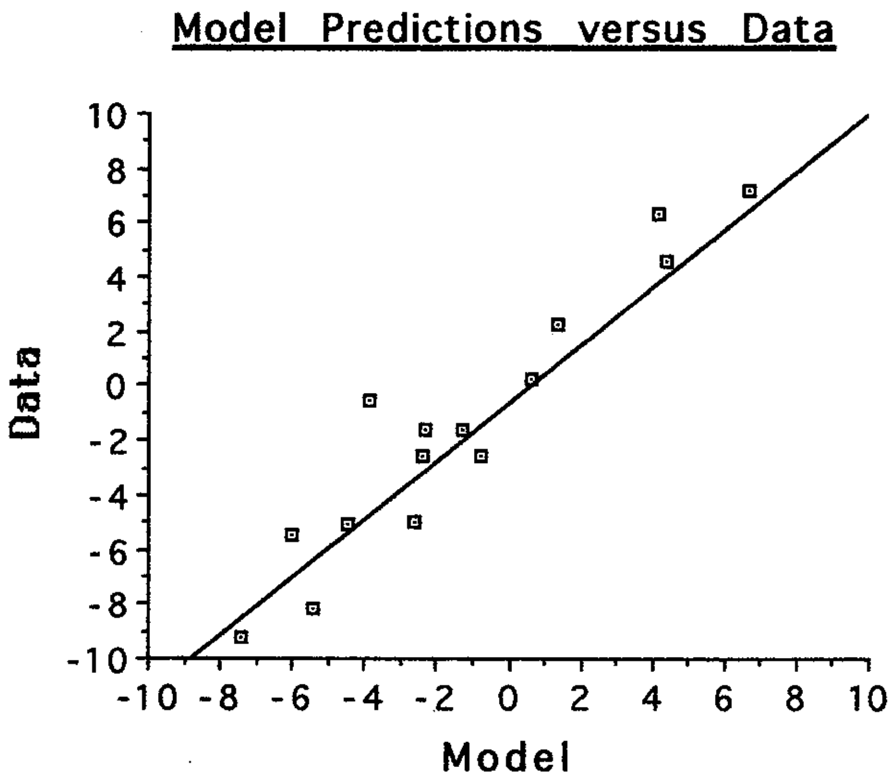
$$d_{ij}(n+1) = S_i(n+1) - S_j(n+1)$$

$$= (1-\Theta)\, d_{ij}(n) + \Theta\, P_{ij}(n)$$

Ordering of Team Strengths




First Experiment
Point-spreads
Game Number
15
10
5
0
-5
-10
-15
0
20
40
60
80
Model
Data


Model Predictions versus Data
10
8
6
4
2
0
-2
-4
-6
-8
-10
Data
-10 -8 -6 -4 -2 0 2 4 6 8 10
Model

Home-Court Advantage (H)

Predicted Point-Spread when I is the home team:

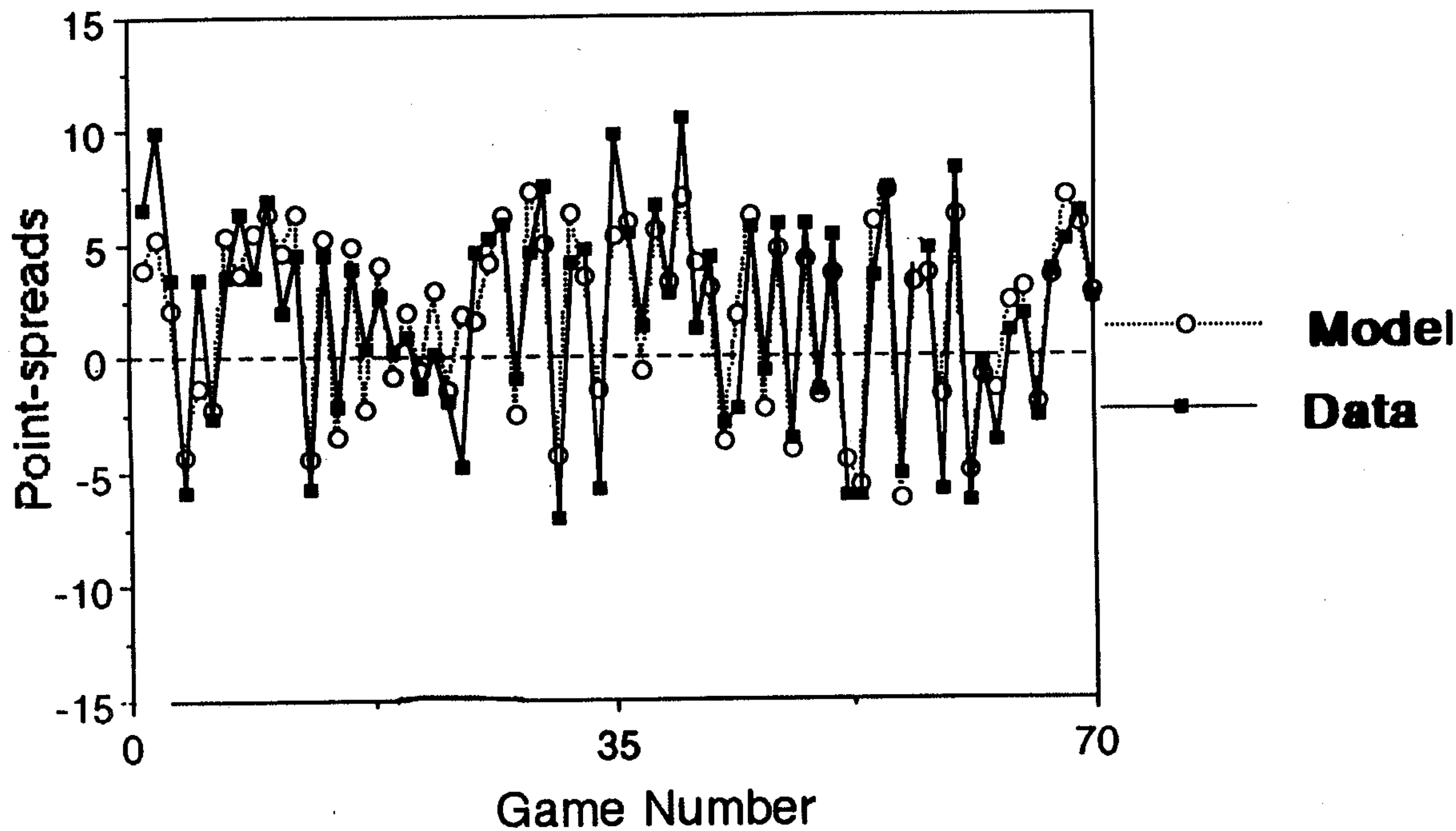
$$\text{Spread}(n) = S_i(n) - S_j(n) + H$$

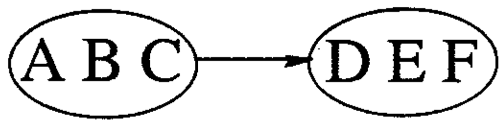
Learning Equations when I beats J by $P_{ij}(n)$ points:

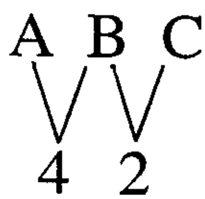
$$\Delta S_i = -\Delta S_j$$

$$= .5\Theta\left[P_{ij}(n) - \text{Spread}(n)\right]$$

Home-Away Experiment



1) 

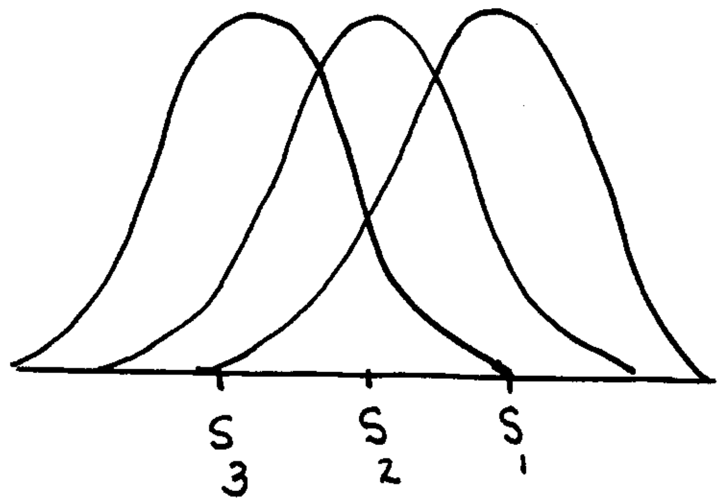
2) 

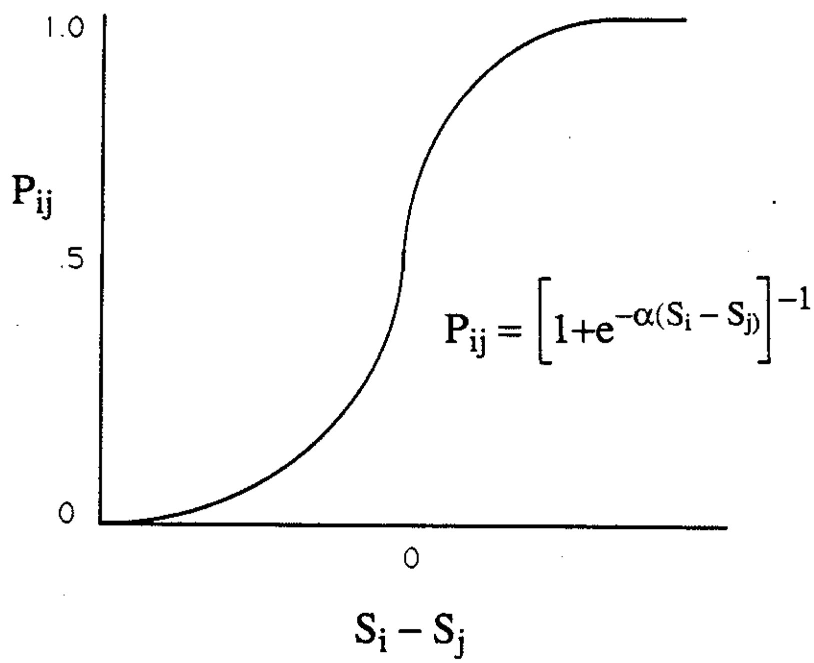
==> A-B=4

A-C=2

A-C=6

3)





4)