ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 66, No. 1, April, pp. 16–21 (1996)
ARTICLE NO. 0034

A Strength Model of Probability Judgments for Tournaments

DEREK J. KOEHLER

Stanford University

Fans of the National Basketball Association (NBA) assigned probability judgments to the outcomes of upcoming NBA games, and rated the strength of each team involved. The probability judgments obtained from these "expert" subjects exhibited high intersubject agreement and also corresponded closely to the eventual game outcomes. A simple model that associates a single strength value with each team accurately accounted for the probability judgments and their relationship to the ratings of team strength. The results show that, in this domain at least, probability judgments can be derived from direct assessments of strength which make no reference to chance or uncertainty. © 1996 Academic Press, Inc.

In a recent article, Tversky and Koehler (1994) proposed a new model of subjective probability, called support theory, in which the judged probability of a hypothesis is given by the support (or strength of evidence) of that hypothesis normalized relative to the support of its alternative. Furthermore, it was suggested that in some situations, probability judgments can be predicted from independent ratings of evidence strength. The present article extends these notions from judgments regarding a single process (e.g., the winner of a horse race) to judgments regarding multiple processes (e.g., the results of a tournament).

Consider a simple tournament in which each of several teams or individuals, denoted A, B, C, plays once against each of the others. Let $P(A > B)$ be the judged probability that A will win its game against B. Assume that there are no ties and that probability judgments satisfy binary complementarity, so that $P(A > B) + P(B > A) = 1$. In general, there is no necessary relationship between the probabilities associated with different matches. For example, judges may assign high values to $P(A > B)$ and to $P(B > C)$, and a low value to $P(A > C)$. Neither standard probability theory nor support theory constrains the relationship among these three estimates. Nevertheless, it is suggested that in many situations people's judgments about the outcomes of a tournament may satisfy the following simple model.

Assume that for each team in the tournament, the judge has a value $s(A)$, interpreted as the strength of team A. The judged probability that team A will beat team B, then, is given by the following *strength model:*

$$P(A > B) = \frac{s(A)}{s(A) + s(B)} \qquad (1)$$

According to this model, the judged probabilities associated with the results of a tournament depend only on the strengths of the respective teams. In other words, the model assumes no interactions; no team is expected to play especially well or especially poorly against any specific opponent.

It is convenient to restate the strength model in terms of the probability ratio $R(A > B) = P(A > B)/P(B > A)$, that is, the odds for A against B. Assuming, for simplicity, that all probabilities are positive, Eq. (1) yields $R(A > B) = s(A)/s(B)$. It is easy to verify that this model implies the following product rule:

$$R(A > B)R(B > C) = R(A > C). \qquad (2)$$

Furthermore, it can be shown that this rule, in conjunction with binary complementarity, is not only necessary but also sufficient for Eq. (1). Indeed, the present model is formally equivalent to the binary version of Luce's (1959) choice model. The only difference is that the latter applies to choice frequencies, whereas the present model applies to probability judgments.

In the current study, the strength model is applied to probability judgments of avid fans of the National Basketball Association (NBA) about the outcomes of upcoming games. To do so, the effect of the home court should be taken into account. This could be accomplished by estimating two strength values for each team, one as a home team and one as a visiting team.

This research was conducted while I was being supported by a National Defense Science and Engineering Graduate Fellowship. I am grateful to Amos Tversky for helpful discussion and to David Freedman for advice regarding statistical analyses used in this study. Address correspondence and reprint requests to Derek J. Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1 Canada.

0749-5978/96 $18.00

There is, however, a more parsimonious model in which the home-court advantage increases the strength of all teams by the same factor.

Let $P(A^* > B)$ denote the judged probability that Team A will beat Team B at A's court, and let $R(A^* > B) = P(A^* > B)/P(B > A^*)$ denote the corresponding odds. Assume that there exists a constant $q > 1$, reflecting the home-court advantage, such that $R(A^* > B) = qs(A)/s(B)$, or equivalently $P(A^* > B) = qs(A)/[q\,s(A) + s(B)]$. It can be shown that this assumption holds if and only if the ratio $R(A^* > B)/R(A > B^*)$ is a constant, independent of A and B. This form is closely related to Luce's (1959) representation of response bias.

The present model accounts for $n(n - 1)$ judgments regarding all matches between n teams (in both locations) in terms of n parameters ($n - 1$ strength values and a home-court factor). This model is compared to more general forms that permit interactions between teams and that allow for a variable home-court advantage. The model is also used to compare the strength measure derived from probability judgments with direct estimates of team strength.

Let $\hat{s}(A)$ be the rated or assessed strength of team A. It is natural to assume that direct assessments of team strength and the strength value derived from judged probability are monotonically related; that is, $\hat{s}(A) \geqslant \hat{s}(B)$ if and only if $s(A) \geqslant s(B)$. Furthermore, it has been suggested (Tversky & Koehler, 1994) that the corresponding ratios are also monotonically related; that is, $\hat{s}(A)/\hat{s}(B) \geqslant \hat{s}(C)/\hat{s}(D)$ if and only if $s(A)/s(B) \geqslant s(C)/s(D)$. It can be shown that if these two conditions hold and both scales are defined, say, on the unit interval, then there exists a constant $k > 0$, such that the two measures of strength are related by a power transformation of the form $s(A) = \hat{s}(A)^k$ (cf. Theorem 2 of Tversky & Koehler, 1994, p. 567).

The following study tests the above assumptions and explores the possibility of predicting probability judgments from direct ratings of team strength that make no reference to chance or uncertainty.

METHOD

Subjects ($N = 90$) were NBA fans who subscribe to a computer bulletin board (newsgroup), called rec.sport.basketball.pro, that is internationally accessible through the Internet. The questionnaire was posted to the newsgroup and readers were encouraged to complete the form and return it via electronic mail to the experimenter within a week. Participants were offered entry in a lottery that gave a one-in-ten chance of winning $15. The study was conducted over a one-week period (February 6–13, 1993) approximately halfway into the 1992/1993 regular NBA season.

TABLE 1

Team Standings at the Time of the Study (February 9, 1993)

Team	Wins	Losses	Winning percentage	Games back
Phoenix	34	9	.791	—
Portland	28	15	.651	6
LA Lakers	24	22	.522	11
Golden State	20	27	.426	16
Sacramento	16	29	.356	19

All participants completed a two-part questionnaire. The first asked them to assess the probability that the home team would win in each of 20 upcoming basketball games. The games were chosen in the following way. Five of the seven teams in the NBA's Pacific Division were selected for study (Phoenix, Portland, Los Angeles Lakers, Golden State, and Sacramento). Only five teams were selected so that all pairwise comparisons could be elicited without requiring a prohibitively long questionnaire. These teams were selected to maximize the range of team strength—both the best and worst teams in the division were included. Table 1 lists the winning records of the teams at the point in the season when the study was conducted.

The 20 outcomes to be evaluated consisted of all possible matches among the five teams, each appearing twice, so that each member of the pair was designated once as the home team and once as the visiting team. For each match, subjects were asked to assess the probability that the home team would win in the next game between the two teams on that team's court. Subjects were instructed that 0% indicated absolute certainty that the visiting team would win, 50% indicated that either team was equally likely to win, 100% indicated absolute certainty that the home team would win, and that intermediate numbers indicated intermediate degrees of certainty. For each game under evaluation, the home team was listed first for easier reading.

In the second part of the questionnaire, subjects rated the strength of each team. They were instructed as follows:

> First choose the team you believe is the strongest of the five, and SET THAT TEAM'S STRENGTH TO 100. Assign the remaining teams strength ratings in proportion to the strength of the strongest team. For example, if you believe a given team is half as strong as the strongest team (the team you gave a 100), give that team a strength rating of 50. Equally strong teams should be given equal strength ratings.

These instructions were intended to prompt subjects to formulate their ratings using a ratio scale.

Subjects returned the completed questionnaire by electronic mail to the experimenter. Any questionnaires received after the one-week deadline were not considered in the analyses that follow.

RESULTS

Of the 90 subjects who completed the questionnaire, two were identified as clear outliers. The correlation between their judgments and the actual game outcomes for these two subjects was +.12 and −.47 (corresponding to approximately 4 and 7 standard deviations below the mean, respectively). These same two subjects also showed very low agreement with other subjects, with correlations between their judgments and the set of mean judgments being only +.31 and −.73 (corresponding to approximately 3 and 8 standard deviations below the mean, respectively). As these two subjects appeared to be outliers in terms of both validity and agreement, their data were dropped from subsequent analyses.

Accuracy

As a preliminary analysis, the quality of subjects' judgments was examined by assessing their relationship to the actual game outcomes. Between the time of the study and the end of the regular basketball season, all of the games considered by subjects had been played except for one (Sacramento at Portland; due to irregularities in the schedule these teams played twice at Portland before the time of the study and thus did not play there again during that season). The outcomes for 19 of the 20 games, then, can be used to assess the accuracy of subjects' probability judgments. Judgments for the game that was not played were ignored in the analyses which follow. The outcomes for the 19 games are listed in Table 2 along with the mean judged probability associated with each game.

As indicated in Table 2, judged probability was an excellent predictor of the winning team in each match. In fact, the home team won 10 of the 12 games in which it was assigned a mean probability greater than .5, and lost all 7 games in which it was assigned a mean probability of less than .5. By this fairly crude analysis, accuracy (as measured by the proportion of times the game outcome was predicted by the direction of the judged probability's deviation from .5) was 89% (17/19) for the mean judgments. A similar analysis of individual subjects' judgments (in which all judgments of exactly .5 were ignored) showed that the median subject accuracy was 86%, and that 67 of the 88 subjects achieved at least 80% accuracy. It is clear, then, that subjects' judgments in this study provided quite accurate predictions of the eventual game outcomes.

TABLE 2

Final Score for Each of the 19 Games Evaluated, Listed in Order of the Mean Judged Probability of a Home-Team Victory

Home		Visitor		
Team	Score	Team	Score	Probability
Phoenix*	130	Sacramento	122	.91
Phoenix*	115	LA Lakers	114	.89
Phoenix*	111	Golden State	100	.89
Portland*	115	Golden State	99	.84
Portland*	105	LA Lakers	103	.76
LA Lakers*	125	Sacramento	107	.69
Phoenix*	129	Portland	111	.68
LA Lakers*	115	Golden State	112	.63
Golden State*	132	Sacramento	105	.61
Sacramento	101	Golden State*	113	.54
Golden State	111	LA Lakers*	117	.53
Portland*	102	Phoenix	97	.51
Sacramento	99	LA Lakers*	104	.49
LA Lakers	105	Portland*	109	.47
Golden State	96	Portland*	113	.36
LA Lakers	105	Phoenix*	120	.34
Sacramento	111	Portland*	113	.32
Golden State	100	Phoenix*	122	.28
Sacramento	108	Phoenix*	128	.21

* Indicates the winning team in each match.

Probability Judgments

In the present experiment, strength can be derived directly from the probability judgments, or it can be measured independently through the strength ratings. The probability judgments are examined first, followed by the strength ratings.

As expected, use of this expert population yielded highly reliable judgments. The means of the 20 probabilities were computed and correlated with each individual's judgments. The median correlation was .93, and 62 of the 88 subjects had correlations of .9 or higher.

Across the 10 pairs of teams, subjects assigned a given team a probability of winning that was on average greater by .177 when the team was playing at home than when it was playing on the opponent's court. This difference corresponded very closely to the actual difference of .188 for the five teams at that point in the season.

Consider first the strength model, which can be used to estimate the strength values from the probability judgments in log-odds form. To derive the log-odds form, recall that the odds ratio $R(A > B) = P(A > B)/P(B > A) = s(A)/s(B)$. Translation to a logarithmic scale yields the linear equation $\log R(A > B) = \log s(A) - \log s(B)$ relating log odds to the strength measure. Because

TABLE 3

Summary of Models Tested and Their Fit to the Probability Judgment Data

Model	Home-court factor	Parameters	R^2 value for group data	Median R^2 value for individual data
Strength	Constant	4 team + 1 home-court	.980	.890
Interaction	Constant	10 interaction + 1 home-court	.987	.934
Strength	Variable	4 team + 5 home-court	.989	.944
Interaction	Variable	10 interaction + 5 home-court	.995	.977

strength is defined up to a ratio scale, the strength of an arbitrarily selected team can be set equal to 1, and the strength values of the remaining four teams can then be estimated from the probability judgments given by each subject.

To account for the home-court advantage, recall that $P(A^* > B)$ denotes the judged probability that Team A will beat Team B at A's court; $R(A^* > B)$ is the corresponding odds. The assumption of a constant home-court advantage $q > 0$, described above, implies that $\log R(A^* > B) = \log s(A) - \log s(B) + Q$ in log odds, where $Q = \log q$.

All analyses were conducted using least-squares multiple regression in the log-odds metric. Analyses were conducted both for the set of mean ratings (over subjects) and separately for each subject. Table 3 provides a summary of the models tested and their fit to the data. The strength model requires estimation of four strength values and the home-court factor. The R^2 value for the analysis of mean data was .980, indicating that the assumption of a single strength value for each team can account for virtually all of the variance in the probability judgments.

The strength model can be compared to a more general *interaction model* in which the probability judgments are assumed to depend on the specific pair of teams involved in each game. The interaction model estimates a value for each possible pair of teams (i.e., each match), and assumes only additivity of complementary pairs. This model allows for the possibility that a team may play above or below its usual level depending on the specific opponent involved. The form of the interaction model is $\log R(A^* > B) = I(A > B) + Q$, where $I(A > B) + I(B > A) = 0$ and Q is the home-court advantage. The interaction model requires estimation of 10 interaction parameters (one for each possible match) and the home-court factor. The R^2 value for the analysis of mean data was .987, which is not significantly greater than that of the strength model, $F(6, 9) = 0.82$. The strength model, then, is sufficient to account for the probability judgments without assuming interactions among teams.

The results of regression analyses conducted separately for each subject are consistent with this conclusion. For the strength model, the median subject had an R^2 value of .890. For the interaction model, the median R^2 was .934. About as many people (6 of the 88 subjects) exhibited significant ($p < .05$) increases in R^2 with the interaction model as would be expected by chance under the null hypothesis. Examination of the residuals in these analyses showed no indication of a non-normal distribution or of a significant correlation between the predicted values and their residuals.

The strength and interaction models can also be compared assuming that the home-court advantage is not constant but instead varies from team to team. Under this assumption the strength model is $\log R(A^* > B) = \log s(A) - \log s(B) + Q(A)$, where $Q(A)$ is the home-team advantage associated with Team A. This model, which requires estimation of four strength values and five home-team factors, yielded an R^2 value of .989 for the set of mean data. The model did not improve significantly on the assumption of a single home-court advantage for all teams, $F(4, 11) = 1.41$, *ns*, suggesting that a single home-court factor is sufficient in this context. The interaction model assuming a different home-team factor for each team is $\log R(A^* > B) = I(A > B) + Q(A)$, where $I(A > B) + I(B > A) = 0$. This model requires estimation of 10 interaction parameters and 5 home-court factors, and yielded an R^2 value of .995 for the mean data. As was the case when a single home-court factor was assumed, the increase in R^2 from the strength model to the interaction model was not significant, $F(6, 5) = 0.95$. Analyses of judgments given by individual subjects produced similar results. In summary, little predictive power was lost by assuming no interactions among teams and a constant home-team advantage over teams.

These analyses show that the judged probability of basketball games can be described in terms of a model which takes into account only a single parameter for each team, and that more general models allowing interactions between teams do not improve upon this model's performance. It should be noted, however, that the strength model of Eq. (1) is not the only possible form in which the judged probability can be expressed

TABLE 4

Mean Strength Ratings and Probability Judgments. Matrix Entries Reflect the Judged Probability that the Row Team Will Beat the Column Team

			Golden	LA	
Strength/Team		Sacramento	State	Lakers	Portland
100	Phoenix	85	80	78	59
85	Portland	75	74	65	
60	LA Lakers	60	55		
50	Golden State	54			
42	Sacramento				

as a separable function of the two teams involved in the game. For example, one logical alternative is the simple linear model $P(A > B) = s(A) - s(B)$, in which the probability judgment itself, rather than the log-odds transformation of the probability judgment, is described as the difference in strength between teams A and B. When supplemented with a home-court factor such that $P(A > B) = s(A) - s(B) + q$, this *subtractive model* fits the present data as well as the strength model. (For the set of mean judgments, $R^2 = .990$; the median R^2 value for the individual subject data is .926.) Further research will be needed to test whether the difference in strength values is better applied to judgments expressed in the probability metric or in the log-odds metric.

Strength Ratings

Once the strength values have been estimated from the probability judgments, the issue of using the direct ratings as an independent measure of team strength can be addressed. Because separate strength ratings were not obtained for each team playing at home and away, in this analysis the probability judgments assigned to one team's beating another are collapsed across the two possible locations of the game. This was accomplished by assuming binary complementarity, that is, by defining $P(B > A^*) = 1 - P(A^* > B)$, and then computing the mean of the two estimates $P(A^* > B)$ and $P(A > B^*)$ that a given team will win. (For evidence supporting the assumption of binary complementarity see, e.g., Wallsten, Budescu, & Zwick, 1992; for a review, see Tversky & Koehler, 1994.) This process yields a set of 10 pairs of complementary judgments. Table 4 lists the mean probability judgments obtained in this process, showing only the more probable member of each complementary pair of judgments. The mean strength rating of each team is also listed.

Analysis of the strength ratings was as follows. Assuming the power transformation $s(A) = \hat{s}(A)^k$ as described in the introduction, the value of k can be estimated directly from the probability judgments by setting $R(A > B) = [\hat{s}(A)/\hat{s}(B)]^k$. Formulating the model in log odds, we have $\log R(A > B) = k[\log \hat{s}(A) - \log \hat{s}(B)]$. In this formulation, then, k is estimated by the slope of the regression line predicting $\log R(A > B)$ from $\log \hat{s}(A) - \log \hat{s}(B)$.

Using this procedure, the value of k was estimated separately for each subject from his or her set of probability judgments and strength ratings. The median value of k was 1.8 and the mean was 2.2. All subjects had a positive k estimate; 79 of the 88 subjects had a k estimate greater than 1; and 70 of the 88 subjects had a k estimate that was less than 3. Once the value of k is estimated, it can be used along with the five strength ratings to fit the probability judgments given by each subject. The median R^2 value for the regression analyses (conducted separately for each subject) was .87. By comparison, the regression analysis using a single k for all subjects yielded an R^2 value of only .63, suggesting that there were individual differences in the use of the strength rating scale.

A similar technique was used to estimate the value of k for the set of 20 mean probability judgments and 5 strength ratings. Here k was estimated to have a value of 1.9, and the resulting R^2 value was .97. Thus using k to transform the strength ratings yields a near-perfect correlation between judged probability and the normalized (transformed) strength ratings.

If Team A is twice as strong as Team B, the value of k obtained in this experiment suggests that the judged odds of A beating B will be close to 4 to 1. One speculation is that the value of k may reflect the relative predictability of the outcome variable in question. Thus, for example, considerably lower values of k would be expected if subjects were asked to judge the probability that the home team will score first in the game (rather than that the home team will win the game) because this variable is generally less predictable. One way to think about k, then, is in terms of sample size. In the example above, in which A is deemed twice as strong as B, the outcome variable can be conceptualized as being determined by drawing a random sample of balls from an urn containing two-thirds A balls and one-third B balls. A game's outcome would correspond to a large sample, in which the odds of a sample with more A than B balls are considerably more extreme than 2 to 1. The prediction of which team will score first would correspond to a substantially smaller sample size.

CONCLUSION

The results of this study demonstrate that probability judgments for tournaments can be represented in terms of the normalized strength of the two teams involved in a given game. A strength model based on an extension of support theory (Tversky & Koehler, 1994) accurately accounted for the probability judgments of basketball fans predicting game outcomes; more general models imposing fewer constraints (e.g., an interaction model or a model using a different home-court advantage parameter for each team) did not improve on the quantitative fit achieved by the strength model. Furthermore, direct ratings of team strength also accounted for the probability judgments. The latter result is particularly interesting because it demonstrates that probability judgments can be derived from direct assessments of evidence that make no reference to uncertainty. This finding is consistent with support theory and also with Griffin and Tversky's (1992) analysis of confidence judgments.

The model of judgment implied by this analysis offers the judge a substantial reduction in the computational complexity of the problem domain. In the case of basketball, only one parameter per team is required to account for the full set of probability judgments. Thus, the judged probability assigned to all $n(n - 1)$ possible games among n teams can be computed using only n parameters. Such a model will necessarily miss any interactions among specific pairs of teams, but considerable research suggests that high accuracy often can be maintained even when such potential interactions are ignored (see, e.g., Dawes, Faust, & Meehl, 1988). Furthermore, because each of the n parameters is determined by a strength assessment, the model can produce a full set of probability judgments using mental assessments that involve no uncertainty and that, consequently, correspond more closely to the way people naturally formulate such judgments—namely, in terms of strength or support rather than in terms of probability.

Received: February 3, 1995

REFERENCES

Dawes, R. M., Faust, D., and Meehl, P. E. (1988). Clinical versus actuarial judgment. *Science,* 243, 1668–1674.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology,* 24, 411–435.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* 101, 547–567.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1992). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science,* 39, 176–190.