The Statistician (1997)
46, *No.* 4, *pp.* 541–550

Measuring the effectiveness of playing strategies at soccer

By RICHARD POLLARD†

California Polytechnic State University, San Luis Obispo, USA

and CHARLES REEP

Torpoint, UK

[Received April 1997. Revised August 1997]

SUMMARY
Using a notational system which records on-the-ball events taking place throughout a soccer match, the game can be broken down into a series of team possessions. To assess the effectiveness of a team possession, a quantitative variable is developed representing the probability of a goal being scored, minus the probability of one being conceded. This variable, called the yield, can be used to evaluate both the expected outcome of a team possession originating in a given situation, as well as the actual outcome of the possession. In this way, the effectiveness of different strategies occurring during the possession can be quantified and compared.

Keywords: Logistic regression; Performance analysis; Soccer; Strategy

1. Introduction

Soccer is undoubtedly the world's most popular team sport. The current World Cup competition is being contested by 174 nations and the Fédération Internationale de Football Association, the international governing body, has more member countries than the United Nations. A winning soccer team will bring great prestige to the country or city that it represents, and the coach and players involved can expect to receive international acclaim, as well as huge financial rewards. It is therefore surprising that there have been so few objective and scientific approaches to the most effective ways of playing the game. Soccer is now a big business, and it is difficult to think of any other business activity in which vital decision-making would be tolerated in the almost total absence of the collection and analysis of numerical data. The traditional method of assessing performance has been for an observer to watch a game take place and then to draw subjective conclusions about individual and team performances. Although video recording now enables subsequent viewing of a match, the planning of tactics, strategies and an overall style of play still remain for most coaches an entirely subjective procedure. The purpose of this paper is to develop a method in which the effectiveness of different tactics and strategies can be quantified and compared.

In contrast with team sports such as cricket, baseball and American football, the fundamental difficulty in attempting to make an objective study of team performance at soccer is a lack of routinely recorded quantitative data. The first requirement is thus a method of obtaining such data. One of us (CR) has, since the early 1950s, been using a notational technique which enables a complete record of a match to be recorded on paper. The continuous action of a game is broken down into a series of discrete on-the-ball events (such as a pass, a centre or a shot) and a detailed categorization made for each type of event, for which shorthand codes have been

†*Address for correspondence*: Statistics Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA.
E-mail: rpollard@calpoly.edu

 0039–0526/97/46541

developed. For example, each pass made in a game is classified and recorded by its length, direction, height and outcome, as well as the positions on the field at which the pass originated and ended. Over 40 years ago, a careful analysis of the recorded data isolated particular patterns of play that were most likely to be associated with the scoring of goals and a distinctive style of play was subsequently formulated. This very direct style has been adopted by a handful of teams with great success, notably Wolverhampton Wanderers in the 1950s and Watford in the early 1980s. However, it has been the subject of intense controversy in the soccer world, well summarized by Pitt (1992), but often leading to hostile criticism and ridicule especially from the media.

Details of this system of analysing performance are not well documented and only limited attempts have been made at a formal statistical analysis of the data produced. In the first paper to link statistics and soccer, Reep and Benjamin (1968) showed how the negative binomial distribution provided a good fit to various events in the game. Several attempts to model goal scoring by Poisson, negative binomial and related distributions have been made and are summarized by Pollard (1993). Different playing styles have been quantified and classified by using performance analysis data (Pollard *et al*., 1988). The quantification of the outcome of different strategies to be described in this paper is based on Pollard (1989). Data recorded from videotapes of 22 matches in the 1986 World Cup finals in Mexico were used for the analysis.

Mention should be made of some other systems of soccer recording that are in use, although there appears to have been almost no use of statistical methodology to analyse the mass of data available. In the early 1980s, a method of directly entering the action of a game onto a small hand-held computer was developed in Canada (Franks *et al*., 1983). Analysis of the data produced has been made and some statistical aspects considered (Franks, 1988). A similar method of recording games followed in England (Church and Hughes, 1987) and other methods are in existence in Finland (Paukku, 1994), the USA (Newsweek, 1994) and doubtless elsewhere. There have been three World Congresses of Science and Football: at Liverpool in 1985, Eindhoven in 1991 and Cardiff in 1995. The abstracts and proceedings from each of these congresses include a section on match analysis and provide a good overview of different attempts that have been made to analyse team performance (Reilly *et al*. (1988, 1993) and in the *Journal of Sports Sciences*, volume 13 (1995), pages 499–522). However only one paper, by Ali (1988), hints at more than a very basic use of statistical methods.

2. Measurement of outcome

The basic unit of measurement on which the analysis was made is called a team possession. A team possession starts when a player gains possession of the ball by any means other than a pass from a player of the same team. The player must have enough control over the ball to be able to have a deliberate influence on its subsequent direction. The team possession may continue with a series of passes between players of the same team but ends immediately that one of the following events occurs.

(a) The ball goes out of play.
(b) The ball touches a player of the opposing team (e.g. by means of a tackle, an intercepted pass or a shot being saved). A momentary touch that does not significantly change the direction of the ball is excluded.
(c) An infringement of the rules takes place (e.g. a player is offside or a foul is committed).

Each team possession consists of several components. For example, it may contain a long forward pass hit from midfield. To assess the effectiveness of these components the outcome of each team possession needs to be quantified. Possible outcome variables and their limitations were considered in the following order.

2.1. *Goals*

The obvious outcome measure for a team possession in soccer is whether or not it results in a goal being scored. However, of nearly 6000 team possessions recorded, only 47 resulted in goals. Thus if goals were to be used as a binary outcome variable over 99% of team possessions would be classified as a failure with the same value, much information would be lost and a larger sample would be needed for a meaningful analysis.

2.2. *Shots*

A goal is normally preceded by a shot, a shot being defined as a direct attempt to score a goal by a player striking the ball at the opponents' goal. Although much less of a rare event than goals (8% of team possessions produced shots) difficulties still exist in the use of shots as an outcome variable. The probability that a shot scores varies greatly with both the location of the shot as well as other quantifiable factors (Olsen, 1988; Pollard, 1995). For example, shots from central locations inside the penalty area are on average over 15 times more likely to produce goals than shots from outside. It was therefore decided that an improved measure of outcome might be attained by assigning each shot a weight, according to its estimated probability of scoring.

2.3. *Weighted shots*

The outcome variable for shots was assigned a value 0 if the team possession failed to produce a shot and a value p if a shot resulted, p being the estimated probability that the shot would score. An estimate for p was derived from the results of a logistic regression analysis based on the team possessions that resulted in a shot. Details of this analysis are described in Section 3. Even with this improved method of dealing with shots, 92% of team possessions continued to have an outcome value of 0. For the team possessions that failed to produce a shot, it seemed desirable to distinguish between possession that had been relatively successful (e.g. produced a corner) compared with possession that had clearly been less successful (e.g. losing possession before moving the ball beyond the halfway line). An outcome variable that reflected such differences was sought.

2.4. *Yield*

For the yield outcome measure, each team possession was first classified by two variables, the zone of origin and the type of possession, defined as follows. The pitch was divided into six zones in accordance with Fig. 1 and the zone in which the team possession originated was recorded. The analysis of shots had suggested that the probability of scoring depended on whether or not the possession originated as a set play (such as a free kick) or in open play, this binary information being represented by 'type of possession'. For possession of type j starting in zone i, the probability p_{ij} of scoring a goal could be estimated by

$$p_{ij} = \sum_{k=1}^{n_{ij}} p_{ijk} / n_{ij}$$

where $i = 1, \ldots, 6$ depending on the zone, $j = 1$ (open play) or $j = 2$ (set play), p_{ijk} denotes the kth team possession of type j originating in zone i and is equal to the estimated scoring probability p of that possession if it ends in a shot or is otherwise 0 (i.e. p_{ijk} represents the weighted shot value and n_{ij} is the total number of team possessions originating in zone i of type j.

Thus $p_{51} = 0.014$ would signify that, of 1000 team possessions originating in zone 5 as open play, teams would expect to score 14 goals.

Having established values for each p_{ij}, it was possible to assign one of these values to each recorded team possession, depending on the outcome of that possession and the zone of origin of the *next* team possession. For example, if the end of a team possession was immediately

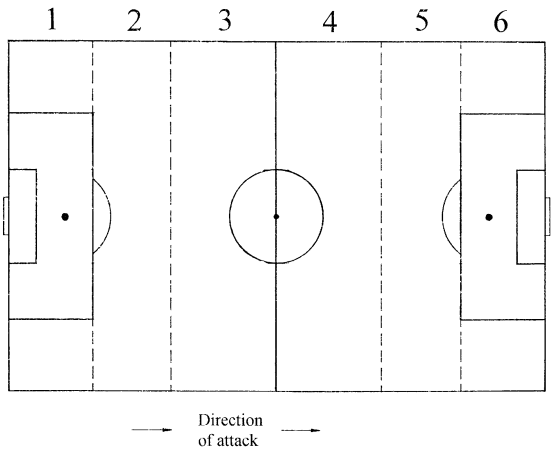

Fig. 1. Division of the field of play into six zones

followed by the same team regaining possession as a set play in their zone 4, then the outcome value for the first team possession would be p_{42}. However, if possession was not regained, but transferred to the opposing team in open play, then the outcome value for the first possession would be $-p_{31}$, since subsequent possession would be with the opposing team in their own zone 3. The negative sign thus indicates that the initial team possession ended with possession for the opponents, and the value p_{ij} is the expected probability that the opponents will score from the situation in which they gained possession.

At this stage the outcome of each team possession could be assigned the value p, p_{ij} or $-p_{ij}$, which we shall refer to as preliminary yields. This compares with the values p or 0 using weighted shots. Using these new values, the average outcome value for team possessions originating in each zone of each type could be recalculated using preliminary yields in place of weighted shots in the formula above. The new average outcome value was called the yield y_{ij} and became the new estimate of the outcome value of a team possession of type j originating in zone i.

Each actual team possession could now be reassigned an outcome value determined, as before, by how and where the subsequent team possession began. This outcome value was the yield as defined above, replacing the preliminary yield. This iterative process was continued so that at each stage new average yield outcome values were calculated based on the yields from the previous iteration. When none of the y_{ij} changed by more than 0.001 at an iteration, the process was terminated and stable yield values established.

Yield became our final outcome variable. Its two distinct uses should be emphasized. Firstly, yield could be used to quantify the *expected* outcome of a team possession of type j originating in zone i. Secondly, yield could be used to measure the *actual* outcome of a team possession, based on the team, zone and type of the subsequent possession.

For example, if the yield $y_{51} = 0.025$, then every 1000 team possessions originating in open play in zone 5 would be expected to produce 25 more goals than would be conceded. Alternatively, a team possession that ended and was followed by the same team regaining possession in open play in zone 5 would be assigned an outcome value of $y_{51} = 0.025$.

In terms of probability, the yield of a team possession is the estimated probability of scoring a goal minus the estimated probability of conceding a goal, based on the outcome of the possession. Although the yield is less easy to interpret than a goal or a shot, it has the major advantage of quantifying and distinguishing between the vast majority of team possessions that fail to produce a shot. In fact the development of the yield has a direct link

with goal scoring. The yield is based on estimated scoring probabilities which are themselves based on the scoring of a goal through the logistic regression analysis.

The main characteristics of the four outcome variables discussed are summarized in Table 1. The calculated values of each y_{ij} are given in Section 4, together with some applications and possible uses to which the yield can be put.

3. Calculation of shot probabilities

3.1. *Method of analysis*

The outcome variable 'weighted shots' required values for the probability of scoring a goal under various circumstances. To estimate these probabilities an analysis was done on the 489 team possessions that resulted in a shot. A goal was scored from 47 of these, giving an average scoring probability of 0.096, which is consistent with a scoring ratio of about 1 in 10 found in many other studies and summarized in Pollard (1995), where the relationship between shots and goals was also investigated in relation to the location of the shot. Fig. 2 displays an

TABLE 1
Comparison of values assigned to different outcome variables†

Outcome of team possession	*Outcome variable*				
	Goal	*Shot*	*Weighted shot*	*Preliminary yield*	*Yield*
Shot: goal	1	1	p	p	p
Shot: not a goal	0	1	p	p	p
Possession regained	0	0	0	p_{ij}	y_{ij}
Possession lost	0	0	0	$-p_{ij}$	$-y_{ij}$

†p is the estimated scoring probability of a shot, p_{ij} is the estimated probability of scoring from possession originating as type j in zone i and y_{ij} is the estimated yield from possession originating as type j in zone i.

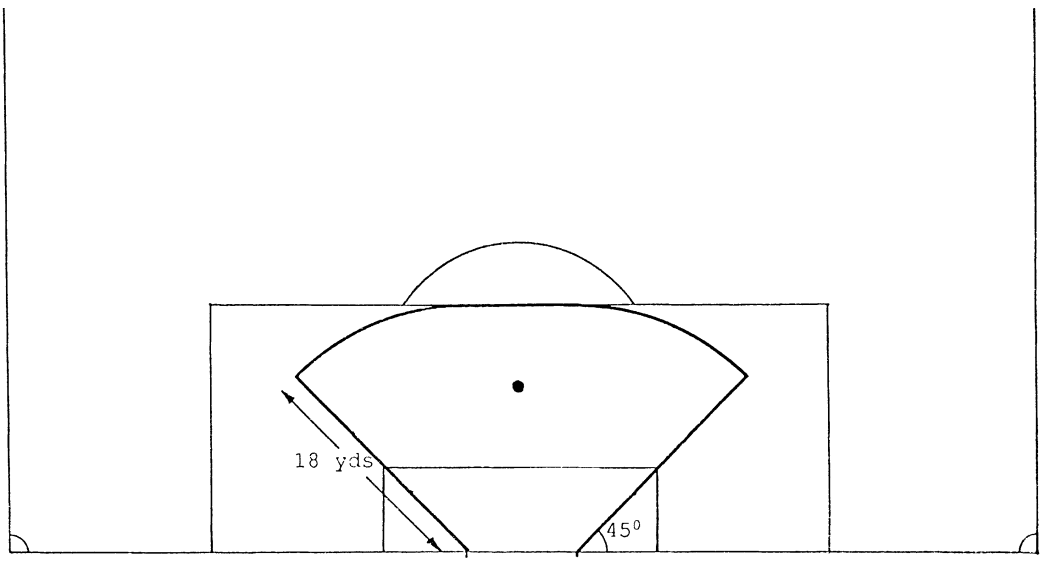

Fig. 2. Location of the arc of relatively high goal scoring probability

arc inside the penalty area from which most goals are scored. The probability of scoring from a shot inside the arc was 0.189 ($n = 206$), compared with 0.014 from outside ($n = 278$). These figures exclude five penalties.

Clearly the location of the shot has a major influence on the probability of scoring. An initial assessment of other factors that are likely to influence goal scoring was carried out on the 489 team possessions that ended in a shot. Three of these factors appeared to be of possible importance and it was therefore decided that a goal scoring model would include the scoring or not of a goal as the dependent variable, with the following explanatory variables: x_1 is the distance in yards between the midpoint between the goalposts and the position of the shot; x_2 is the angle in radians of the arc formed by the intersection of a line from the position of the shot to the nearest goalpost, and from the same goalpost parallel to the side-line (0 if the position of the shot is directly in front of the goal); $x_3 = 0$ if the player taking the shot touched the ball only once or $x_3 = 1$ if there was one touch or more before the shot; $x_4 = 0$ if the player taking the shot was less than 1 yard from the nearest defender or $x_4 = 1$ if the player was more than 1 yard away; $x_5 = 0$ if team possession originated from open play or $x_5 = 1$ if possession originated from a set play.

With a binary dependent variable and a mixture of binary and continuous explanatory variables, logistic regression was used. This enabled the probability of scoring a goal (p) to be estimated from any combination of the explanatory variables and, if necessary, their interactions. The analysis was performed separately for kicked and headed shots since the effect of the explanatory variables would probably be different for each type of shot. The regressions were developed as a series of generalized linear models by using the statistical package GLIM (Payne, 1985).

3.2. *Kicked shots*

No interaction was significant nor was the variable x_3 which represented the number of touches of the ball. The final model, based on an analysis of 410 kicked shots, was

$$y = 1.245 - 0.219x_1 - 1.578x_2 + 0.947x_4 - 1.069x_5$$

from which the probability of scoring could be estimated as

$$p = \frac{\exp y}{1 + \exp y}.$$

Thus the outcome for each team possession resulting in a shot could be represented by the probability of scoring, calculated by substituting the observed values of the four explanatory variables into the above equations. These were the values used for the weighted shots described in Section 2.3. For example, consider a shot from 16 yards, directly in front of goal, with an opponent within 1 yard and from a team possession originating as a set play. The value of y is -3.328 and hence the probability of scoring estimated by $p = 0.035$.

The model lends itself to other interpretations. For example, from the coefficient of x_1, $\exp(0.219) = 1.24$ indicates that every yard nearer goal increases the probability of scoring by 24%. Similarly, $\exp(0.947) = 2.58$ suggests that a player who manages to be over 1 yard from an opponent when shooting more than doubles his probability of scoring.

An alternative way of displaying the model is to construct contours of scoring probability for various situations. For example, for a kicked shot in open play, less than 1 yard from the nearest opponent, Fig. 3 shows a selection of scoring probability contours.

3.3. *Headed shots*

A similar analysis was carried out for the smaller sample of 74 headed shots. Since only one of these resulted from more than one touch of the ball, variable x_3 was not included. No

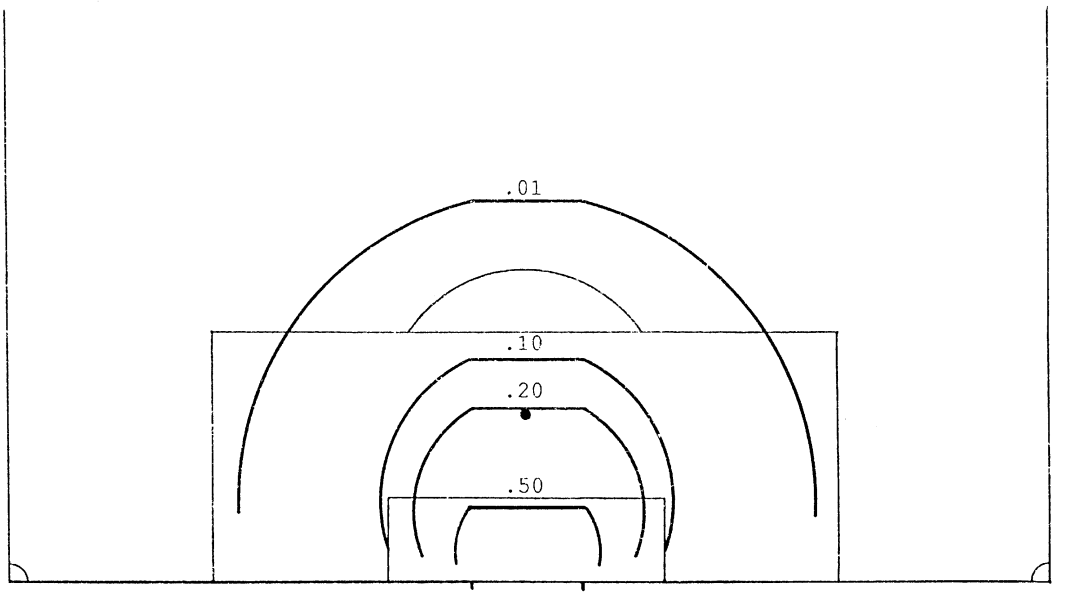

Fig. 3. Scoring probability contours for kicked shots from open play and from less than 1 yard from the nearest opponent

interaction was significant; nor was variable x_4, representing the distance from the nearest opponent. The final model was

$$y = 1.520 - 0.237x_1 - 3.117x_2 - 1.784x_5.$$

This enabled scoring probabilities to be estimated in the same way as for kicked shots.

3.4. *Penalty kicks*

Only five penalty kicks were recorded, but previous studies have estimated the probability of scoring at 0.79 in North America (Pollard, 1986) and 0.81 worldwide (Pollard, 1995). The goal scoring probability for penalties was therefore estimated as 0.8.

3.5. *Applications*

For each of the 489 team possessions that resulted in shots, the probability of scoring was estimated by entering details of the shot into the appropriate model. This became the 'weighted shot' outcome value described in Section 2.3 and on which the yield, the final outcome variable, was based.

4. Evaluation of expected outcome

In Section 2.4, the concept of the yield as a measure of the outcome of a team possession was introduced. The yield is based on the scoring probabilities which were developed in Section 3. If used to estimate the expected outcome of a team possession, the yield will depend on both the zone of origin of the team possession in addition to the type of play (open play or set play). These expected yield values are shown in Table 2 and are based on the 5844 team possessions recorded from the 1986 World Cup in Mexico.

The yield is easier to interpret when given as a rate per 1000 team possessions. The values

TABLE 2
Yield per 1000 team possessions, classified by zone of origin and type of play

Zone of origin	*No. of team possessions for the following types of play:*		*Yield for the following types of play:*	
	Open play	*Set play*	*Open play*	*Set play*
1	865	651	5.9	2.2
2	822	244	8.5	0.5
3	837	321	6.2	2.2
4	473	450	10.9	8.5
5	318	336	24.8	12.6
6	111	416	78.3	18.0

then give the net yield in goals scored per 1000 team possessions in each situation. For example, the yield of 24.8 in Table 2 signifies that for every 1000 team possessions originating as open play in zone 5 a team can expect to score 24.8 more goals than it concedes.

For all zones, team possessions originating in open play had a higher yield than those starting as set plays. This is presumably a reflection of the extra time that a set play gives the opposing team to move into position to defend. The very high yield in zone 6 from open play (78.3) suggests a net yield of one goal for about every 13 such possessions. This clearly indicates the importance that teams should attach to regaining possession of the ball in this zone, constituting positions in and around the opponents' penalty area. This is one of the fundamental strategies of a direct and aggressive style of play, an assessment of which inspired a journalist to describe the 1982 Watford forwards as playing like 'wild dogs hurling themselves at a brick wall'.

It is perhaps surprising that the yield, the outcome variable of a team possession, should be so dependent on the type of possession and the zone of origin, both of which are initial components of the possession. It is indeed quite possible that as a team possession develops the effect of these initial components becomes less important. However, the majority of team possessions in soccer are of very short duration. In a series of over 23000 team possessions analysed in 1958 in the English first division, fewer than 5% consisted of four or more completed passes (Reep *et al.*, 1971). Even in the 1986 World Cup games under analysis, where teams were in general striving to keep possession, the corresponding figure was barely 15%. It is for this reason that the average yield values in Table 2 are still found to be so sensitive to the initial components of the team possession.

5. Application to playing strategy

The yield can also be used to quantify the *actual* outcome of a team possession as described in Section 2.4. Thus, to assess the effectiveness of a particular strategy, the mean yield of all team possessions in which the strategy was used can be calculated. Different strategies can be subsequently compared. As a simple example, consider a team that has a throw-in level with the opponents' penalty area, which means in zone 6. There are two basic strategies; a long throw towards the goalmouth or a short throw to a player of the same team to continue the attack. Expressing the outcome as the yield per 1000 team possessions, in the games recorded in the 1986 World Cup, 32 long throws had a yield of 21.7, compared with 98 short throws with a yield of 3.5. Using a nonparametric test (the distribution of yield values was definitely not normal), this sixfold difference was clearly significant ($P < 0.01$).

Table 3 gives the yield of a selection of different situations and strategies whose effectiveness can be quantified in this way. A negative yield value for an event means that the

TABLE 3
Yield per 1000 team possessions from playing strategies in different situations

Situation	*Strategy*	*n*	*Yield*
Goal kick	Long	99	−2.7
Throw-in in own half	Short	276	−0.2
Possession in zone 4	Short passing only	1372	11.1
	Running with the ball	288	16.3
	Long forward pass	148	23.1
Free kick in zone 5	Direct shot	60	12.5
	Other	143	16.8
Throw-in in zone 6	Short	98	3.5
	Long towards goalmouth	32	21.7
Centres from zone 6	Above waist height	240	33.3
	Below waist height	103	96.6

event would result, on average, in more goals being conceded than scored. It should be emphasized that the data from which this table was constructed was based on a sample of matches from the 1986 World Cup in Mexico. This involved national teams whose playing styles will have been influenced by the stage of the competition and also by the playing conditions (e.g. high altitude and very hot temperature). Nevertheless it is clear that the methodology developed in this paper can highlight substantial differences in the effectiveness of different strategies.

6. Conclusion

With a few notable exceptions, our experience has been that soccer coaches, players, fans and the media are deeply sceptical and often suspicious, to the point of paranoia, at the suggestion that a statistician might have something useful to offer in the way of tactical analysis. Perhaps it is understandable that the preservation of a soccer mystique is something that those professionally involved in the game would want to maintain. Nevertheless, because of the overwhelming importance of winning, anything that might give a coach an additional advantage might at least be expected to be worthy of investigation. Table 3 gives a glimpse of a few of the several hundred different situations and strategies for which a measure of yield, and hence effectiveness, is now available. Although these values relate specifically to the 1986 World Cup, it would not be difficult to repeat the analysis for another competition or for a specific team. A coach could then be pointed towards strategies of high yield, as well as being able to measure objectively the performance of one particular team in terms of its effectiveness in applying any given strategy, old or new. The reliable quantification of the outcome of playing strategies has the potential to be of great use to anyone concerned with the planning of both an overall playing style as well as specific tactics. While developing the methodology for the analysis, somewhat crude descriptions of what constitutes a strategy have been used. In future work, it is hoped to make more refined definitions for different strategies and also to consider further the statistical aspects of the measure of yield so that, for example, confidence intervals can be given.

References

Ali, A. H. (1988) A statistical analysis of tactical movement patterns in soccer. In *Science and Football* (eds T. Reilly, A. Lees, K. Davids and W. J. Murphy), pp. 302–308. London: Spon.

Church, S. and Hughes, M. (1987) A computerized approach to soccer notation analysis. In *Abstr. 1st Wrld Congr. Science and Football, Liverpool*. p. 20. Liverpool: Liverpool Polytechnic.

Franks, I. M. (1988) Analysis of association football. *Soccer J.*, **33**, no. 5, 35–43.

Franks, I. M., Goodman, D. and Miller, G. (1983) Analysis of performance: qualitative or quantitative. In *Science Periodical on Research and Technology in Sport*, GY-1, Mar. Ottawa: Coaching Association of Canada.

Newsweek (1994) The Americans' secret weapon. *Newsweek*, July 11th, 41.

Olsen, E. (1988) An analysis of goalscoring strategies in the World Championship in Mexico, 1986. In *Science and Football* (eds T. Reilly, A. Lees, K. Davids and W. J. Murphy), pp. 373–376. London: Spon.

Paukku, T. (1994) And it's another goal . . .. *New Scient.* June 11th, 30–32.

Payne, C. D. (ed.) (1985) *The GLIM System Release 3.77 Manual*. Oxford: Numerical Algorithms Group.

Pitt, N. (1992) Has football in England lost its way? *Sunday Times*, Oct. 4th, 12.

Pollard, R. (1986) Soccer performance analysis and its application to shots at goal. *Res. Bi-A. Movemnt*, **4**, 19–27.

———(1989) A statistical evaluation of team performance at soccer. *Doctoral Dissertation*. University of the South Pacific, Suva.

———(1993) The random nature of events at football. In *Proc. 22nd Spring Conf. Bulgarian Mathematicians*, pp. 187–193. Sofia: Union of Bulgarian Mathematicians.

———(1995) Do long shots pay off? *Soccer J.*, **40**, no. 3, 41–43.

Pollard, R., Reep, C. and Hartley, S. (1988) The quantitative comparison of playing styles at soccer. In *Science and Football* (eds T. Reilly, A. Lees, K. Davids and W. J. Murphy), pp. 309–315. London: Spon.

Reep, C. and Benjamin, B. (1968) Skill and chance in association football. *J. R. Statist. Soc.* A, **131**, 581–585.

Reep, C., Pollard, R. and Benjamin, B. (1971) Skill and chance in ball games. *J. R. Statist. Soc.* A, **134**, 623–629.

Reilly, T., Clarys, J. and Stibbe, A. (eds) (1993) *Science and Football II*. London: Spon.

Reilly, T., Lees, A., Davids, K. and Murphy, W. J. (eds) (1988) *Science and Football*. London: Spon.