Paper : Game Analytics

Achievement Standards and Causal Structure of Offensive Skill Measurement Items in Japan Professional Football League

Hirotaka Jo[*,], Hiroki Matsuoka[***], Kozue Ando[***] and Takahiko Nishijima[***]**

[*]Department of Sports Science, Shizuoka Sangyo University
1572-1 Oowara, Iwata-city, Shizuoka-pref 438-0043 Japan
[**] Doctoral Program in Health & Sport Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8574 Japan
[***] Faculty of Health & Sport Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8574 Japan
jotk.lab@gmail.com

[Received November 4, 2021; Accepted January 21, 2022]

Since it is not easy for analysts in the instruction field to process the performance big data of soccer games, it is necessary to develop models and analysis methods using big data and return them to the instruction field. Game performance data is measured by a quantitative scale, but if it is converted into achievement type binary data, the measurement cost is reduced, and a criterion-referenced evaluation scale with IRT applied can be constructed. However, converting to binary data reduces the amount of information, so it is necessary to reconfirm whether it reflects soccer skills. Therefore, this study's purpose was to clarify the causal structure of game performance big data with binary data in soccer. To that end, we analyzed the achievement standards for constructing binary data using quantitative game performance big data in J.League offensive plays. The achievement standards were calculated by CART using the Gini impurity, and achievement data converted into a binary scale was constructed. As a result of applying the maximum likelihood method and factor analysis of oblimin rotation to the achievement data and clarifying the factor structure, seven subfactors of offensive skill were extracted. As a result of a path analysis from the variance-covariance matrix of the factor scores estimated by the maximum likelihood method, the causal structure that reflects the attacking style of soccer was clarified in the achievement data. The factor structure and causal structure were superior to the quantitative data.

Keywords: football, offensive play, measurement items, achievement standards, causal structure

[Football Science Vol.19, 59-77, 2022]

1. Introduction

Sports analytics began with quantitative evaluation by frequency in the 1950s and went through quantitative evaluation of game situations by expert's assessment, as well as physical and technical evaluations of passes, running distance, sprints, etc. Nowadays, it is mainly performed through dynamic tactical evaluation by patterns, combinations, interactions, and complex KPIs (Memmert and Raabe, 2019). Because of the development of measuring equipment and software in recent years, the quantity and quality of performance data in many ball games like soccer are improving, and that data is now called "big data in sports" (Dmonte and Dmello, 2017).

It is said that "big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more" (Viktor and Kenneth, 2013). Although this statement is mainly aimed at the economy and government, big data in sports has the potential to provide new insights in various scenarios such as game analysis, development of tactics and strategies, technical guidance, and scouting and training of athletes.

However, without certain information-processing skills, it is not easy for a sports analyst on the competition fields to manage and analyze big data.

Also, while there are many articles about big data in sports, academic discussions are very limited, which points to the need for more studies in this field (Nicholas et al., 2021). To this end, it is necessary to promote studies that use big data in sports to develop theories, models, and analysis and management methods that are profitable at the academic level, and then return the profits to the sports fields.

One of the research themes designed to meet such demand is the development of a soccer skills evaluation scale using game performance data (Jo et al., 2014; Matsuoka et al., 2020). There is a growing number of studies on the composition of skill evaluation scales applying the item response theory (IRT), and they have explored criterion-referenced evaluation standards for running exercises (Aoyagi, 2002), pitching form (Aoyagi, 2006), discus throwing motion skills (Ono et al., 2014), and skill evaluation on baseball (Abe et al., 2017). Studies on soccer include the computerized adaptive testing (CAT) of tactical skills (Ando et al., 2018a; 2018b), two-dimensional evaluation scale of technique and tactics (Jo et al., 2014), and criterion-referenced evaluation items of defensive skills (Matsuoka et al., 2020).

IRT has been attracting attention because of some advantages that are not found in the classical test theory (Ohtomo, 1996) and for enabling criterion-referenced evaluation. However, the correct answer rate in tests (correct/wrong answer pattern) originates from the Rasch model (Rasch, 1960), which is calculated with a function of the product of the answerer's ability and item difficulty. Because of this, interval scale data cannot be used in IRT, and it is necessary to use binary data indicating success or failure or staged reaction data that maintains an ordered relationship.

Binary data or staged reaction data should be used according to the study’s purpose and the field’s characteristics. The former can be applied to “achieved” and “not achieved” phenomena in sports, which makes it very practical and convenient for on-site measurement and suitable for measuring exercise capacity (Aoyagi, 2005). Also, Ono et al. (2014) and Matsuoka et al. (2020) applied the IRT of a 2-parameter logistic model (2PLM) to binary performance data and created criterion-referenced evaluation criteria from estimated item characteristics. Therefore, it can be said that an academic mathematical model has already been incorporated into sports instruction fields. These suggest that in skill evaluation using soccer game performance data, it may be worth using binary data that measure the achievement level based on the criterion-referenced evaluation.

However, since soccer game performance data measure quantitative data, such as the distance covered and the number of passes, it requires the process of conversion into binary data. This problem can be solved with a method that determines the value of achievement standards from a classification binary tree analysis of each item (Jo et al., 2014; Matsuoka, et al, 2020). However, since nominal scales generally have less information than interval scales, when this type of conversion is applied, the characteristics of the data may change.

If the items meet the prerequisites of IRT — being one-dimensional — the items used look as though they are measuring specific skills. However, because IRT does not assume subfactors, it cannot be developed into a causal structure analysis. Causal structure analysis plays an important role in studies of soccer skills. That is because once the causal structure is identified, it is possible to analyze the construct validity of the subfactors that comprise soccer skills, which enables the skill evaluation at a subfactor-level and understanding of skill characteristics (Suzuki, et al., 2000; Yamada et al., 2000). In other words, apart from one-dimensionality in IRT, it is highly significant to perform a factor analysis on achievement binary data assuming soccer skill subfactors and analyze the causal structure.

Previous research that analyzed the subfactor structure of soccer skills includes a study that used ordinal scale data of small scale (Suzuki and Nishijima, 2002; Oe et al., 2007) and another that used quantitative game performance data of the Japan Professional Soccer League (hereinafter “J.League”) (Jo et al., 2022b). However, no previous study has looked at large-scale game performance data from J.League to analyze the causal structure of skill factors using achievement binary data.

Based on the above, this study sought to verify whether achievement binary data from soccer games are appropriate for measuring soccer attacking skills. To this end, it aims to clarify the causal structure of attacking skills from binary performance big data. Also, as a preliminary process, we looked at large-scale game performance data of offensive plays in J.League and analyzed the achievement standards to construct achievement binary data. If this study’s purpose is achieved, it will be possible to collect data with a simple measurement of whether the achievement standards of the measurement items were met or not. This will clarify whether attacking skills can be measured even if the data is binary. If so, it will be possible to apply IRT to game performance data of offensive plays. The two hypotheses below were analyzed in this study:

Hypothesis 1: In J.League's big data of offensive plays, an achievement standard is formed for the 44 items involved in soccer attacking performance.

Hypothesis 2: Achievement binary data has causal relationships between attacking skill subfactors and factors involved in soccer attacking performance, and the factor structure and causal structure are valid and compatible.

2. Method

2.1. Definition of terms

The main terms used in this study are defined as follows:

(1) Play data

Each line is a data set representing a play, which was defined according to Jo et al. (2022a) as the period between the moment a player gets the ball and when he loses it, shoots it, or receives a foul. Then, we counted the number of items contained in multiple motions within a play, based on each item's definition, and this count was set as the measured value of the play data. For example, if a certain play is composed of six motions — pass, trap, dribble, pass, trap, and shoot — the measured value of the item number of "passes" is 2.

(2) Attacking skills

Since the play data is an aggregation of motions of multiple players (e.g., dribble, pass), the factor to be measured is the group attacking skills in soccer.

(3) Measurement item "Shoot"

One of the measurement items on the play data. It represents whether each play had a shot or not with 1 and 0. The same definition applies to the variable "Shoot."

(4) Achievement standard

A threshold for identifying high or low attacking skills defined for each measurement item. Each play is classified as "achieved" or "unachieved" according to this threshold.

(5) Achievement data

It refers to the data set of all items of the play data converted to binary — achieved or unachieved — based on the achievement standards. This type of data is usually called "Pass/Failure type" or "Success/Failure type." However, this study considers the possibility of applying it to criterion-referenced evaluation (Ando et al., 2018a; Matsuoka et al., 2020) and calls it "achievement data."

(6) Game performance data

It refers to a data set composed of the performance directly measured from the motions performed by the players during a soccer match. The play data and achievement data analyzed in this study are included in the game performance data.

(7) Measurement items and variables

"Measurement items" in the field of metrology and "variables" in the fields of statistics and data analysis have similar properties. This study uses one or the other according to the context, but unless specified, these are considered synonymous.

2.2 Study procedure

This study was conducted as follows:

(1) J.League's offensive play data were prepared according to the measurement items by Jo et al. (2022a).
(2) Based on the procedure proposed by Jo et al. (2014) and Matsuoka et al. (2020), the achievement standard of each item was analyzed by classification binary tree analysis with "Shoot" as a dependent variable.
(3) The play data were converted into achievement data based on the achievement standards.
(4) Exploratory and confirmatory factor analyses were performed using the achievement data. The same analyses were performed on the play data for comparison.
(5) Using a logistic regression analysis with "Shoot" in the achievement data as a dependent variable along with a path analysis between attacking skills subfactors, the causal structure in the attacking skills subfactors was analyzed. The path analysis of the play data was analyzed for comparison.

2.3. Analysis targets

This study used the game performance data of offensive plays in all J.League matches played in 2011. This is ball touch data that registers a line every time the player with the ball makes motions such as passing and dribbling. Since it was measured by a staff member that received a certain amount of training by Data Stadium Inc., it is reliable. In 2011, 686 J.League matches were carried out between Division 1 (hereinafter “J1”) and Division 2 (hereinafter “J2”), and 1,312,117 lines of offensive game performance data were measured during that year.

To create the play data, we processed this data according to the definition of attacking skills measurement items proposed by Jo et al. (2022a). However, since set plays are sometimes executed with a special attacking pattern, plays that begin with corner kicks, penalty kicks, and free kicks (both direct and indirect) were excluded. The final play data contained 147,302 plays ×45 items (including the dependent variable “Shoot”). The definition of some play data items and basic statistics are indicated in **Table 1**.

2.4. Analysis method

2.4.1 Achievement standard

Based on the procedure proposed by Jo et al. (2014) and Matsuoka et al. (2020), we performed classification binary tree analysis by CART (Breiman et al., 1984) using Gini impurity (GI) as a branching index on each item. The formula of Gini impurity became as follows:

$$GI = 1 - \sum_{j=0}^{1} \left(\frac{n_j}{n}\right)^2$$

Here, j is the standard of the objective variable (a standard that takes 0 or 1 indicating whether the player took the shoot or not), n is the total number of plays, and n_j is the number of plays that belong to standard j.

Next, the formula for information gain (IG) is shown below. IG in this study indicates how a shoot’s presence or absence can be identified after branching.

$$IG = GI_p - (GI_a P_a + GI_b P_b)$$

Here, GI_p is the Gini impurity before branching, GI_a and P_a are the Gini impurity of plays with a value equal to or higher than the branch value and its percentage, and GI_b and P_b are the Gini impurity of plays with a value smaller than the branch value and its percentage. P_a+P_b equals 1. Based on this formula, the branch value was increased repetitively, from the minimum to the maximum value of the predictive variables, to obtain the information gain. When doing so, the lower limit of the frequency of nodes after branching was set to 1% of the total (n=1,470).

After everything was calculated, the branch value when the information gain showed the maximum value was defined as the achievement standard of that item. To judge the plays as “achieved” or “unachieved,” they were classified into those with a value equal or higher than the branch value and those with a smaller value. Then, the respective shooting rate was calculated, and those with a higher shooting rate were considered “achieved,” and the lower as “unachieved.” This process was performed on all predictive variables to build the achievement data where 1 means “achieved” and 0 means “unachieved.”

2.4.2 Exploratory factor analysis

If the play data is converted to achievement data, the amount of information decreases, which may change the characteristics of the data. Also, it is relevant to identify the causal structure between subfactors in skill evaluation scales, so an exploratory factor analysis was made using the achievement data. Exploratory factor analyses can be made with various rotation methods and factor extraction methods, each with different features. However, because it is difficult to use the most appropriate option according to the characteristics of the data in hand, actual analyses are often conducted by trying various rotation and factor extraction methods and finding the most suitable one (Matsuo and Nakamura, 2002). Moreover, without trying all patterns, it is difficult to verify whether the rotation and factor extraction methods of choice found an optimal factor structure.

To solve this problem, Jo et al. (2022b) proposed a function that they named PAHFA (Program that Analyzes all Hyperparameters in Factor Analysis). It treats the rotation methods, factor extraction methods, and the number of factors in factor analyses as hyperparameters, like in machine learning tuning, and makes a factor analysis with all combinations and returns the optimal model. With this function PAHFA, there is no need to find the three elements above by trial and error, and the analyst can focus on

Table 1 Measurement items, the definitions, and the basic statistics in play data

Item no.	Measurement items	Definitions	Data type	Mean	S.D.	Max.	Median	Min.	Number of yes	Number of no	Freq-uency
1	Shoot	Whether shot or not in one play.	Binary						12,346	134,686	147,032
2	Dribble	The number of dribbles in one play.	Integer	0.10	0.30	3	0	0			147,032
3	Pass	The number of passes (including fail) in one play.	Integer	3.47	3.30	44	2	0			147,032
4	Success of pass	The number of successful passes in one play.	Integer	2.68	3.16	41	2	0			147,032
5	Success rate of pass	The rate of successful passes in one play.	Decimals	62.07	35.84	100.00	66.70	0.00			147,032
6	Direct pass	The number of direct pass (pass action with no trapping) in one play.	Integer	0.55	1.01	16	0	0			147,032
7	Consecutive direct pass	The maximum number of consecutive direct pass in one play.	Integer	0.44	0.72	7	0	0			147,032
8	Through-ball	The number of successful through-ball in one play.	Integer	0.09	0.30	3	0	0			147,032
9	Success of through-ball	The number of successful through-balls in one play.	Integer	0.05	0.22	3	0	0			147,032
10	Cross	The number of driving a cross ball in one play.	Integer	0.11	0.33	4	0	0			147,032
11	Success of cross	The number of successful cross-balls in one play.	Integer	0.02	0.16	3	0	0			147,032
12	Trap	The number of trapping (receive the ball and set it down at foot) a ball in one play.	Integer	2.11	2.56	32	1	0			147,032
13	Rebound-ball	The number of getting a rebound ball in one play.	Integer	0.19	0.43	6	0	0			147,032
14	Flick-on	The number of flick-on (touch the ball lightly to direction shift) in one play.	Integer	0.03	0.17	2	0	0			147,032
15	Throw-in	Whether the play started with throw-in.	Binary						37,127	109,905	147,032
16	Feed	Whether the play started with feeding a ball from goalkeeper.	Integer	0.09	0.28	1	0	0			147,032
17	Total of attack action	The number of attack actions in one play.	Integer	6.06	5.81	78	4	0			147,032
18	Duration of attack	The time (seconds) from the start to the end of the attack.	Integer	10.43	10.79	128	7	0			147,032
19	Average time of attack action	The average time (seconds) of attack action in one play. Formula is " = duration of attack / total of attack action".	Decimals	1.66	1.00	19.00	1.56	0.00			147,032
20	Number of attackers	The number of attackers involved in one play.	Integer	3.24	1.93	11	3	0			147,032
21	Total distance	The sum of the distance (meter) between an action and the next action in one play, excluding defensive actions.	Decimals	55.14	64.95	826.03	30.94	0.00			147,032
22	Total vertical distance	The sum of the vertical distance (meter) between an action and the next action in one play, excluding defensive actions.	Decimals	34.65	38.25	499.83	21.67	0.00			147,032
23	Total horizontal distance	The sum of the horizontal distance (meter) between an action and the next action in one play, excluding defensive actions.	Decimals	34.90	46.28	604.33	16.83	0.00			147,032
24	Distance	The distance (meter) between the first action and the last action in one play, excluding defensive actions.	Decimals	27.72	22.01	114.70	22.14	0.00			147,032
25	Vertical distance	The vertical distance (meter) between the first action and the last action in one play, excluding defensive actions.	Decimals	16.28	23.74	102.17	11.33	-93.50			147,032
26	Horizontal distance	The horizontal distance (meter) between the first action and the last action in one play, excluding defensive actions.	Decimals	14.37	14.75	68.50	9.34	0.00			147,032
27	Maximum vertical distance	The maximum value (meter) in each vertical distances in one play.	Decimals	31.56	28.31	104.67	25.83	0.00			147,032
28	Maximum horizontal distance	The maximum value (meter) in each horizontal distances in one play.	Decimals	29.37	24.83	69.33	26.83	0.00			147,032
29	Mean of pass distance	The mean value of pass distances (meter) in one play.	Decimals	13.02	10.01	92.17	13.04	0.00			147,032
30	Standard deviation of pass distance	The standard deviation of pass distances (meter) in one play. Large SD means attack players use short passes and long	Decimals	4.34	5.84	58.87	1.06	0.00			147,032
31	Area of attack	The sum of triangle areas (sq. meter) created by consecutive three actions in one play, excluding defense action.	Decimals	1211.3	1992.0	28700.5	262.00	0.00			147,032
32	Forward propulsion	Set forward (0 degrees) to 1, rightward (90 degrees) to 0, backward (180 degrees) to -1, leftward (270 degrees) to 0, and converted the angles of ball moving into value between -1 and +1. The forward propulsion is the total of the values.	Decimals	12.77	18.93	118.71	8.14	-74.36			147,032
33	Wide propulsion	Set rightward (90 degrees) and leftward (270 degrees) to 1, forward (0 degrees) and backward (180 degrees) to 0, and converted the angles of ball moving into value between -1 and +1. The wide propulsion is the total of the values.	Decimals	27.68	37.69	499.61	12.90	0.00			147,032
34	Proportion of forward	The percentage of forward moving actions in one play excluding defensive actions. The forward moving action is defined as the angle between 315 degrees and 45 degrees.	Decimals	37.23	34.21	100.00	33.33	0.00			147,032
35	Proportion of backward	The percentage of backward moving actions in one play excluding defensive actions. The backward moving action is defined as the angle between 135 degrees and 225 degrees.	Decimals	13.93	22.93	100.00	0.00	0.00			147,032
36	Proportion of rightward	The percentage of rightward moving actions in one play excluding defensive actions. The rightward moving action is defined as the angle between 45 degrees and 135 degrees.	Decimals	20.12	26.62	100.00	6.25	0.00			147,032
37	Proportion of leftward	The percentage of leftward moving actions in one play excluding defensive actions. The leftward moving action is defined as the angle between 225 degrees and 315 degrees.	Decimals	19.66	26.65	100.00	0.00	0.00			147,032
38	Trun back	The total of actions that the angle (internal angle) formed by three consecutive actions is under 60 degrees. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions.	Integer	0.84	1.44	21	0	0			147,032
39	Change in direction	The total of actions that the angle (internal angle) formed by three consecutive actions is over 60 degrees and under 120 degrees. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions.	Integer	1.57	2.43	36	1	0			147,032
40	Same direction	The total of actions that the angle (internal angle) formed by three consecutive actions is over 120 degrees and 180 degrees or less. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions.	Integer	1.33	2.18	30	0	0			147,032
41	Twice speed	The number of actions that moved the ball more than twice as fast as the previous action.	Integer	0.09	0.32	4	0	0			147,032
42	Penalty area	Whether penetrated into penalty area in one play.	Binary						11,042	135,990	147,032
43	Side of penalty area	Whether penetrated into side of penalty area in one play.	Binary						15,498	131,534	147,032
44	30m line	Whether penetrated into 30m area from goal line in one play.	Binary						35,766	111,266	147,032
45	Vital area	Whether penetrated into vital area in one play.	Binary						26,887	120,145	147,032

the selection of the variables. The function PAHFA uses the fa function of the psych package (Revelle, 2020) in R (R Core Team, 2020) to run a factor analysis. Also, it searches for the best adoption criterion in a factor loading matrix of convergent solutions; then, it returns the models, giving priority to those with a high percentage of variables in which the factor loading exceeding that adoption criterion is recognized in only a single factor (simple structure index).

In this study, we implemented the function PAHFA in R version 4.0.1. and tested a total of 80 patterns: five types of factor extraction methods — promax rotation, oblimin rotation, simplimax rotation, geomin rotation, and cluster rotation (each factor extraction method has the methods of maximum likelihood and weighted least squares) and eight factors, from 3 to 10. Then, we discussed the best model returned by the function PAHFA and repeated this process — of selecting the variable and running the function PAHFA — until the construct validity was satisfied as a subfactor of soccer attacking skills. The construct validity was discussed and determined by a licensed soccer coach with more than ten years of experience as an instructor and two university faculty members specializing in soccer data analysis.

We applied the process above to the achievement data and play data and compared which would show a more valid factor structure. For the factor analysis of the achievement data, a tetracholic correlation matrix was used, and for the play data, a correlation matrix that combines a product-correlation coefficient, a polyserial correlation coefficient, and a tetracholic correlation coefficient depending on the variable scale.

2.4.3 Confirmatory factor analysis

Even if a valid factor structure is found by exploratory factor analysis, it is not necessarily compatible with mathematical models. For this reason, we carried out a confirmatory factor analysis on the factor structures found with achievement data and play data. The goodness of fit indices used were GFI, AGFI, CFI, NFI, RMSEA, AIC, χ^2 value, and p-value.

The goodness of fit of a confirmatory factor analysis is usually increased by making a model modification based on a modification index. The goodness of fit can be maximized by modifying the model endlessly, but the result is not necessarily a consistent model that suits the analyst's hypothesis. Therefore, it is necessary to modify the model according to the analyst's hypotheses, previous research, and substantial scientific evidence (Toyoda, 2012).

The purpose of the confirmatory factor analysis of this study is to verify whether the factor structure of attacking skills found by exploratory factor analysis fits the mathematical model, and determining the goodness of fit is not a priority. Therefore, we analyzed the model before the index-based modification and a model that was modified until it was no longer possible to calculate index candidates. The factor structure of the play data was also analyzed to be compared with the achievement data. The software used was the CFA function of the lavaan package in R (Yves, 2012), specifying the same correlation matrix, factor extraction method, rotation method, and the number of factors as those of the model adopted in the exploratory factor analysis.

2.4.4 Causal structure of attacking skills subfactor

The rotation methods specified in the function PAHFA in exploratory factor analysis are both oblique, assuming a correlation between the factors. Since subfactors of attacking skills are likely to have a mutual relationship, instead of functioning individually, it is relevant to identify the relationship between the factors. Structural equation modeling (SEM) is often used in analyses with this kind of objective, but SEM requires a multivariate normal distribution, which conflicts with the achievement data of this study. Therefore, the parameter estimation method must be chosen carefully. Two methods for analyzing binary data are the path analysis between factor scores using the maximum likelihood method (Mitsunaga et al., 2005) and SEM using weighted least squares of asymptotically distribution-free (ADF) (Finney and DiStefano, 2006).

With the former model, the factor score in the measurement equation is calculated with the maximum likelihood estimation of a parametric model, and the sample variance-covariance matrix and factor average vector of the factor score are estimated. Based on those, and by making a path analysis of the structural equation part by maximum likelihood estimation, it is possible to obtain a correct estimation value. Moreover, it is proven that the higher number of samples, the closer it gets to the true value, which makes it suitable for this study that involves a large number of samples. The latter method can output the correct estimation value with a few thousand samples, even if the data does not follow a multivariate normal distribution, so it can be used in this analysis as well.

The results of the exploratory factor analysis of this study showed that the most appropriate factor structure was found in the model using the maximum likelihood method. For this reason, we decided to unify the estimation methods and used the path analysis between factor scores using the maximum likelihood method to analyze the causal structures. The factor scores were calculated by designating "tenBerge" as the argument score in the fa function of the psych package in R. The tenBerge method can calculate factor scores by finding the weights that maintain the correlation between the factors of the oblique solution (Revelle, 2020). To build the path diagram, a soccer player with experience in the Japanese national team joined the group of the soccer coach and university faculty members mentioned before. They carried out the task while checking if the content was appropriate to ensure that the model reflected an attacking style in soccer.

In identifying the causal structure of attacking skills, there is another element that cannot be ignored. This is because the measurement items of this study were used in the shoot prediction model presented by Jo et al. (2022a), it is necessary to verify whether the factors found affect whether a shot was taken or not. However, since the variable "Shoot" is a single dependent variable, it was not used in the factor analysis and, therefore, cannot be included in the variance-covariance matrix between factor scores. Thus, before the path analysis, we used logistic regression with the factor score as an independent variable to analyze which factors have a strong effect on "Shoot." In other words, we identified the factors that affect "Shoot" the most by logistic regression analysis and made a path analysis with those factors as endogenous variables. Finally, we identified the overall causal structure of attacking skills factors by a two-stage analysis.

In the play data of quantitative scale, we also made a path analysis between factors using the variance-covariance matrix between factor scores and compared the differences with the causal structure of the achievement data. Since it is a quantitative scale, SEM could be applied. However, to avoid possible misunderstandings over the presence or absence of a measurement equation (when comparing it with the path diagram of the achievement data), we prioritized clarity and opted for path analysis. In all analyses of this study, the significance level was set to under 0.1%.

2.5. Ethical Considerations

The data for this study were purchased from Data Stadium Inc., and usage permission was obtained. In addition, this study was conducted with the approval of the Research Ethics Committee at the Faculty of Health and Sport Sciences, University of Tsukuba (Approval No.: 30-28).

3. Results

3.1. Achievement standards

The achievement standards were calculated with "Shoot" as the target variable, and the solution of all items converged (**Table 2**). The items with a high shooting rate when the achievement standard is met are "Success of cross" 70.5%, "Penalty area" 56.3%, "Vital area" 40.0%, "Success of through-ball" 35.1%, and "30m line" 26.6%. Therefore, items related to the penetration of the penalty area are at the top, followed by forward movement-related items such as "Vertical distance" 22.9% and "Forward propulsion" 22.5%. On the other hand, the shooting rate when "Proportion of backward," "Throw-in," "Feed," and "Flick-on" were achieved was less than 9.1%, which is almost the same as the shooting rate of all plays (8.4%).

In binary measurement items, if the achievement standard is "=1," it means Yes; if it is "=0," it means No. For example, if the achievement standard of "Vital area" is "=1," if the player penetrates a vital area, it is judged "achieved;" if not, "unachieved." For the measurement items Integer and Decimals, the achievement standards were set with logical equations "equal to or greater than" (>=) or "less than" (<). Based on these achievement standards, the play data was converted into achievement data.

3.2. Exploratory factor analysis

The selection of variables was repeated based on the content of items included in the same factor, as well as the magnitude of the factor loading and commonality. As a result, a simple structure that

Table 2 Achievement standards of attack skill measurement items

Item No.	Item name	Data type	Gini impurity of objective variable	Information gain	Achievement standard	Gini impurity After separation	Shot play rate in achieved node	Number of shot play in achieved node	Number of play in achieved node
2	Dribble	Integer	0.154	0.00361	>= 1	0.340	0.217	2,956	13,623
3	Pass	Integer	0.154	0.00179	>= 4	0.220	0.126	6,291	50,067
4	Success of pass	Integer	0.154	0.00408	>= 3	0.247	0.145	7,600	52,595
5	Success rate of pass	Decimals	0.154	0.01130	>= 92.5	0.327	0.206	8,335	40,473
6	Direct pass	Integer	0.154	0.00193	>= 1	0.224	0.128	6,218	48,502
7	Consecutive direct pass	Integer	0.154	0.00193	>= 1	0.224	0.128	6,218	48,502
8	Through-ball	Integer	0.154	0.00232	>= 1	0.314	0.195	2,457	12,576
9	Success of through-ball	Integer	0.154	0.00663	>= 1	0.456	0.351	2,293	6,534
10	Cross	Integer	0.154	0.00436	>= 1	0.349	0.225	3,270	14,545
11	Success of cross	Integer	0.154	0.01851	>= 1	0.416	0.705	2,428	3,442
12	Trap	Integer	0.154	0.00363	>= 2	0.229	0.132	8,563	64,954
13	Rebound-ball	Integer	0.154	0.00022	>= 1	0.191	0.107	2,694	25,165
14	Flick-on	Integer	0.154	0.00002	= 0	0.155	0.084	12,076	142,930
15	Throw-in	Binary	0.154	0.00003	= 0	0.158	0.086	9,494	109,905
16	Feed	Integer	0.154	0.00007	< 1	0.157	0.086	11,537	134,436
17	Total of attack action	Integer	0.154	0.00672	>= 5	0.252	0.148	9,806	66,293
18	Duration of attack	Integer	0.154	0.00714	>= 10	0.268	0.159	9,039	56,726
19	Average time of attack action	Decimals	0.154	0.00251	>= 1.5	0.202	0.114	9,735	85,318
20	Number of attackers	Integer	0.154	0.00561	>= 4	0.264	0.156	8,023	51,345
21	Total distance	Decimals	0.154	0.00714	>= 42.5	0.262	0.155	9,450	60,999
22	Total vertical distance	Decimals	0.154	0.00634	>= 37.5	0.272	0.162	8,133	50,116
23	Total horizontal distance	Decimals	0.154	0.00707	>= 25.5	0.265	0.157	9,165	58,217
24	Distance	Decimals	0.154	0.00808	>= 31.5	0.279	0.168	9,011	53,724
25	Vertical distance	Decimals	0.154	0.00883	>= 39	0.353	0.229	5,837	25,484
26	Horizontal distance	Decimals	0.154	0.00259	>= 13	0.223	0.128	7,547	58,984
27	Maximum vertical distance	Decimals	0.154	0.00414	>= 30.5	0.232	0.134	8,899	66,320
28	Maximum horizontal distance	Decimals	0.154	0.00306	>= 10	0.198	0.111	10,987	98,652
29	Mean of pass distance	Decimals	0.154	0.00228	>= 10.5	0.196	0.110	10,054	91,035
30	Standard deviation of pass distance	Decimals	0.154	0.00415	>= 2.5	0.231	0.133	8,998	67,423
31	Area of attack	Decimals	0.154	0.00719	>= 593	0.264	0.157	9,321	59,465
32	Forward propulsion	Decimals	0.154	0.00790	>= 31.14	0.349	0.225	5,462	24,222
33	Wide propulsion	Decimals	0.154	0.00694	>= 19	0.263	0.156	9,233	59,351
34	Proportion of forward	Decimals	0.154	0.00275	>= 8	0.193	0.108	11,110	102,481
35	Proportion of backward	Decimals	0.154	0.00086	< 46.5	0.165	0.091	12,011	132,085
36	Proportion of rightward	Decimals	0.154	0.00315	>= 3.5	0.217	0.124	9,096	73,571
37	Proportion of leftward	Decimals	0.154	0.00282	>= 0.5	0.214	0.122	8,871	72,772
38	Trun back	Integer	0.154	0.00156	>= 1	0.208	0.118	6,990	59,260
39	Change in direction	Integer	0.154	0.00531	>= 2	0.265	0.157	7,670	48,840
40	Same direction	Integer	0.154	0.00451	>= 1	0.232	0.134	9,376	70,118
41	Twice speed	Integer	0.154	0.00073	>= 1	0.249	0.146	1,866	12,797
42	Penalty area	Binary	0.154	0.03724	= 1	0.492	0.563	6,215	11,042
43	Side of penalty area	Binary	0.154	0.00262	= 1	0.307	0.189	2,934	15,498
44	30m line	Binary	0.154	0.02126	= 1	0.390	0.266	9,508	35,766
45	Vital area	Binary	0.154	0.04457	= 1	0.480	0.400	10,742	26,887

satisfies the construct validity was recognized in the achievement data when the function PAHFA was executed for the 12th time (**Table 3**).

Because the first factor includes “Forward propulsion,” “Vertical distance,” and “Total vertical distance,” it was named “Forward moving.” The second factor includes “Trap,” “Same direction,” and “Change in direction.” Because these are items that switch from horizontal to vertical movement and vice-versa, or carry the ball in the same direction, it was named “Connection movement.” The third factor includes “Penalty area” and “Vital area,” which refer to the penetration into important areas for shooting, so it was named “Final area.” The fourth factor includes “Cross” and “Side of penalty area,” which represent movements of crossing the ball from the side, so it was named “Side attack.” The fifth factor includes “Maximum vertical distance” and “Maximum horizontal distance,” which represent the maximum distance covered by the ball, so it was named “Width and depth.” The sixth factor includes “Direct pass” and “Pass,” which are items related to passes, a basic motion in soccer, so it was named “Pass.” Lastly, the seventh factor includes “Horizontal distance,” “Wide propulsion,” and “Total horizontal distance,” so it was named the “Sideward moving” factor. These factors show a clear and simple structure with a factor loading of 0.48 as the adoption criterion, and the cumulative contribution rate up to factor seven was 91%.

Meanwhile, the exploratory factor analysis using the play data with 10 repetitions of the analysis generated results close to those of the factor structure of the achievement data (**Table 4**). At this stage, the factors Sideward moving, Connection movement, Side attack, Final area, and Pass were composed of

Table 3 Result of exploratory factor analysis using achievement data

Item no.	Item name	F1	F2	F3	F4	F5	F6	F7	com
F1	**Forward moving**								
32	Forward propulsion	**0.97**	-0.01	0.02	0.03	0.03	-0.02	-0.02	1.00
25	Vertical distance	**0.95**	-0.02	0.02	0.03	0.03	-0.04	0.04	0.95
22	Total vertical distance	**0.64**	0.20	0.00	0.02	0.04	0.24	0.10	1.00
F2	**Connection movement**								
12	Trap	0.04	**0.91**	0.05	0.04	0.05	-0.04	0.05	1.00
40	Same direction	0.13	**0.62**	0.08	0.07	0.00	0.10	0.08	0.81
39	Change in direction	0.06	**0.53**	0.08	0.06	0.08	0.24	0.13	0.88
F3	**Final area**								
42	Penalty area	0.01	-0.11	**0.94**	0.12	0.01	-0.02	0.04	1.00
45	Vital area	0.02	0.18	**0.87**	-0.09	0.02	0.04	-0.04	0.83
F4	**Side attack**								
10	Cross	0.01	-0.03	0.07	**0.96**	-0.03	0.00	0.05	1.00
43	Side of penalty area	0.04	0.05	-0.05	**0.81**	0.05	-0.01	-0.12	0.66
F5	**Width and depth**								
27	Maximum vertical distance	0.14	-0.03	0.04	0.01	**0.95**	0.00	-0.11	1.00
28	Maximum horizontal distance	-0.14	0.14	0.02	0.08	**0.70**	0.07	0.36	1.00
F6	**Pass**								
6	Direct pass	0.02	-0.08	0.06	-0.02	0.03	**0.87**	0.02	0.75
3	Pass	0.01	0.40	-0.05	0.16	0.07	**0.56**	0.05	0.99
F7	**Sideward moving**								
26	Horizontal distance	0.09	0.03	0.05	-0.01	0.02	0.00	**0.67**	0.58
33	Wide propulsion	0.16	0.21	0.12	0.04	0.05	0.20	**0.50**	1.00
23	Total horizontal distance	0.21	0.21	0.09	0.05	0.05	0.19	**0.49**	1.00
	SS loadings	2.93	2.86	2.01	1.92	1.81	1.99	1.91	
	Proportion Var	0.17	0.17	0.12	0.11	0.11	0.12	0.11	
	Cumulative Variance	0.17	0.34	0.46	0.57	0.68	0.80	0.91	
Correlation matrix between factors	F1	1.00	0.50	0.52	0.51	0.54	0.40	0.32	
	F2	0.50	1.00	0.36	0.37	0.44	0.66	0.67	
	F3	0.52	0.36	1.00	0.56	0.41	0.31	0.38	
	F4	0.51	0.37	0.56	1.00	0.40	0.30	0.27	
	F5	0.54	0.44	0.41	0.40	1.00	0.42	0.38	
	F6	0.40	0.66	0.31	0.30	0.42	1.00	0.59	
	F7	0.32	0.67	0.38	0.27	0.38	0.59	1.00	

*This exploratory factor analysis used maximum likelihood method and oblimin rotation.
*Factor loadings over 0.48 were adopted for the simple structure.

the same items as in the achievement data. However, the loading of “Total vertical distance,” “Maximum vertical distance,” and “Maximum horizontal distance” was small with all factors. As a result, the factor “Forward moving” has one less item than the factor structure of the achievement data, and the factor “Width and depth” was not extracted.

The three play data items that were not included in any factor were excluded, and the exploratory factor analysis with the function PAHFA was carried out again. As a result, the factor structure was largely lost, as 8 out of 14 items were included in the first factor, and four factors were extracted. Then, the items were selected again, the items that had been removed were inserted back, and the analysis was repeated, but no appropriate factor structure not extracted. Therefore, we concluded that the results shown in **Table 4** are the factor structure — although not a simple one — in the play data.

3.3. Confirmatory factor analysis

To verify whether the factor structure found by exploratory factor analysis fits a mathematical model, the same factor structure was reproduced and subjected to confirmatory factor analysis. With the

Table 4 Result of exploratory factor analysis using play data

Item No.	Item name	F1	F2	F3	F4	F5	F6	F7	com
F1	**Sideward moving**								
33	Wide propulsion	**0.95**	0.05	0.00	-0.01	0.00	0.01	-0.01	1.00
23	Total horizontal distance	**0.91**	0.07	0.01	-0.01	0.01	0.03	-0.01	1.00
26	Horizontal distance	**0.59**	-0.15	-0.01	0.06	0.01	-0.01	0.03	0.22
F2	**Connection movement**								
12	Trap	0.07	**0.93**	0.03	0.02	0.05	-0.04	0.01	1.00
40	Same direction	0.20	**0.65**	0.06	0.00	0.01	0.08	-0.02	0.82
39	Change in direction	0.23	**0.52**	0.03	0.02	0.09	0.17	0.00	0.83
F3	**Forward moving**								
32	Forward propulsion	-0.02	-0.01	**1.00**	0.02	-0.01	-0.01	0.01	0.98
25	Vertical distance	0.00	0.02	**0.91**	-0.02	0.02	0.00	-0.03	0.83
F4	**Side attack**								
43	Side of penalty area	0.03	0.00	0.07	**0.99**	-0.03	0.01	0.19	1.00
10	Cross	-0.02	0.04	-0.04	**0.84**	0.10	0.00	-0.33	1.00
F5	**Final area**								
45	Vital area	-0.03	0.10	0.00	-0.03	**0.96**	0.04	0.14	0.95
42	Penalty area	0.10	-0.11	0.12	0.11	**0.78**	-0.01	-0.27	1.00
F6	**Pass**								
6	Direct pass	0.09	-0.08	0.02	-0.02	0.05	**0.77**	0.00	0.65
3	Pass	0.04	0.51	0.02	0.08	-0.01	**0.53**	-0.02	1.00
F7	**Undifined factor**								
27	Maximum vertical distance	0.13	0.01	0.13	0.12	0.28	0.03	0.25	0.31
28	Maximum horizontal distance	0.35	0.02	-0.09	0.15	0.17	0.07	0.22	0.32
22	Total vertical distance	0.30	0.29	0.26	0.06	0.08	0.21	0.08	0.88
	SS loadings	3.15	2.69	2.15	1.93	1.99	1.47	0.38	
	Proportion Var	0.19	0.16	0.13	0.11	0.12	0.09	0.02	
	Cumulative Variance	0.19	0.34	0.47	0.58	0.70	0.79	0.81	
	F1	1.00	0.87	0.41	0.28	0.44	0.73	0.03	
	F2	0.87	1.00	0.33	0.20	0.30	0.65	0.12	
	F3	0.41	0.33	1.00	0.36	0.54	0.27	0.03	
	F4	0.28	0.20	0.36	1.00	0.43	0.17	-0.18	
	F5	0.44	0.30	0.54	0.43	1.00	0.29	-0.18	
	F6	0.73	0.65	0.27	0.17	0.29	1.00	0.07	
	F7	0.03	0.12	0.03	-0.18	-0.18	0.07	1.00	

*This exploratory factor analysis used maximum likelihood method and oblimin rotation.
*Factor loadings over 0.52 were adopted for the simple structure.
*In the continuous analysis excluding the 7th factor, the simple structure was never appeared, thus the 7th factor was thus the 7th factor was left as an undefined factor in this analysis.

7-factor model of the achievement data, the GFI, CFI, and NFI before the modification exceeded 0.9, the RMSEA was below 0.1, but AGFI was 0.861. Meanwhile, with the 6-factor model of the play data, the GFI, CFI, and NFI before the modification exceeded 0.9, but the RMSEA was 0.110, and AGFI was 0.858 (**Table 5**). When the model's consistency was ignored, and the goodness of fit increased to the maximum based on the modification indicator, the GFI, AGFI, CFI, and NFI of both the achievement, and play data exceeded 0.99, and the RMSEA was below 0.01. Also, the χ^2 test of the achievement data after the model modification was not significant.

The path coefficient results indicate that "Pass" in the play data before the model modification, as well as "Total vertical distance" in the achievement data and "Pass" in the play data after the model modification, were inappropriate solutions that exceeded 1.0. The path coefficient of the other items was between 0.4 and 1.0 and, therefore, acceptable.

3.4. Causal structure of attacking skills subfactors

3.4.1. Logistic regression analysis

Due to the characteristics of the analysis of this study, performing a path analysis with the seven factors of attacking skills extracted in the exploratory

Table 5 Results of confirmatory factor analysis of seven factors model in achievement data and six factors model in play data

			Achivement data (7 factors model)		Play data (6 factors model)	
			Before modification	After modification	Before modification	After modification
Goodness of fit indices		GFI	0.911	0.999	0.916	0.999
		AGFI	0.861	0.999	0.858	0.999
		CFI	0.912	0.999	0.957	0.999
		NFI	0.912	0.999	0.957	0.999
		RMSEA	0.099	0.002	0.110	0.009
		AIC	140,953.2	297.3	1,107,520.1	233.5
		CHI-SQ	140,843.2	21.3	110,666.1	27.5
		DF	98	15	62	2
		P-Value	<0.001	0.126	<0.001	<0.001
Path coeficients	**Forward moving**					
	Forward propulsion		0.91	0.57	0.97	0.95
	Vertical distance		0.89	0.58	0.93	0.95
	Total vertical distance		0.69	1.05	-	-
	Connection movement					
	Trap		0.83	0.82	0.97	0.95
	Same direction		0.75	0.75	0.90	0.90
	Change in direction		0.81	0.83	0.90	0.93
	Final area					
	Penalty area		0.69	0.59	0.70	0.57
	Vital area		0.76	0.90	0.75	0.92
	Side attack					
	Cross		0.84	0.86	0.85	0.85
	Side of penalty area		0.61	0.59	0.60	0.60
	Width and depth					
	Maximum vertical distance		0.64	0.60	-	-
	Maximum horizontal distance		0.84	0.91	-	-
	Pass					
	Direct pass		0.94	0.92	0.67	0.70
	Pass		0.59	0.60	1.08	1.02
	Sideward moving					
	Horizontal distance		0.53	0.54	0.46	0.46
	Wide propulsion		0.98	0.98	1.00	1.00
	Total horizontal distance		0.98	0.98	1.00	1.00

factor analysis and the measurement item "Shoot" mixed was not an ideal statistics analysis procedure. Therefore, we performed a logistic regression analysis with "Shoot" as a dependent variable and the factor scores of the seven factors as independent variables (**Table 6**). The results showed that "Final area" had the strongest effect on "Shoot" (partial regression coefficient = 1.407, odds ratio = 4.085, $p<0.001$). It was followed by "Connection movement" (partial regression coefficient = 0.347, odds ratio = 1.415, $p<0.001$) and "Forward moving" (partial regression coefficient = 0.143, odds ratio = 1.153, $p <0.001$). The odds ratio results show that the "Final area" was 2.9 times larger than the second-highest factor, "Connection movement." The Nagelkerke R^2, which indicates the goodness of fit of the regression model, was 0.446.

3.4.2. Path analysis

Next, we considered the variance-covariance matrix of factor scores as an observed variable and performed a path analysis by the maximum likelihood method. Since the logistic regression revealed that the penetration into the final area had a strong effect on

Table 6 Results of logistic regression analysis for the dependent variable "shot" using the variance-covariance matrix of factor scores as an independent variable, in achievement data

Factor name	partial regression coefficient	Standard error	z-value of Wald test	p-value	Odds ratio
Final area	1.407	0.012	115.9	<0.001	4.085
Connection movement	0.347	0.018	19.3	<0.001	1.415
Forward moving	0.143	0.014	10.4	<0.001	1.153
Width and depth	0.023	0.017	1.4	0.175	1.024
Sideward moving	-0.024	0.017	-1.4	0.155	0.976
Pass	-0.156	0.015	-10.2	<0.001	0.855
Side attack	-0.516	0.013	-40.6	<0.001	0.597
(Intercept)	-3.238	0.016	-205.1	<0.001	0.039
Nagelkerke R^2	0.446				

whether the player would shoot or not, we designated “Final area” as the final endogenous variable. In addition, we created two causal structure models that reflect attacking styles in soccer. The first is the “Sideward to Forward (SF) Model,” in which the team launches a horizontal offensive and then attacks vertically, and the second is the “Forward to Sideward (FS) Model,” in which the team launches a vertical offensive and then attacks horizontally. The path diagrams of the achievement data were created with seven factors, and the path diagrams of the play data with six factors, and all four were analyzed.

The goodness of fit indices GFI, AGFI, CFI, and NFI exceeded 0.9 with all models. The RMSEA was under 0.1 with both models of the achievement data and under 0.05 with both models of the play data.

In the SF model of the achievement data (model SF-A) shown in **Figure 1**, no path coefficient exceeded 0.6 and, overall, the coefficients were low to moderate. The paths with a moderate effect were “Pass → Sideward moving” at 0.59, “Pass → Connection movement” at 0.39, “Sideward moving → Connection movement” at 0.44, “Sideward moving → Width and depth” at 0.43, “Connection movement → Width and depth” at 0.34, “Connection movement → Forward moving” at 0.35, “Width and depth → Forward moving” at 0.40, “Forward moving → Side attack” at 0.40, and “Side attack → Final area” at 0.38.

In the FS model of the achievement data (model FS-A), shown in **Figure 2**, no path coefficient exceeded 0.55 and, like the case above, the path coefficients were low to moderate, overall. The paths with a moderate effect were “Pass → Forward moving” at 0.40, “Pass → Connection movement” at 0.54, “Connection movement → Sideward moving” at 0.51, “Connection movement → Width and depth” at 0.34, “Forward moving → Width and depth” at 0.32, “Forward moving → Side attack” at 0.40, and “Side attack → Final area” at 0.38.

In the SF model of the play data (model SF-P), shown in **Figure 3**, a strong path coefficient was identified in “Pass → Sideward moving” at 0.73 and “Sideward moving → Connection movement” at 0.86. Also, a moderate path coefficient was identified in “Sideward moving → Forward moving” at 0.57, “Sideward moving → Side attack” at 0.30, and “Forward moving → Final area” at 0.35.

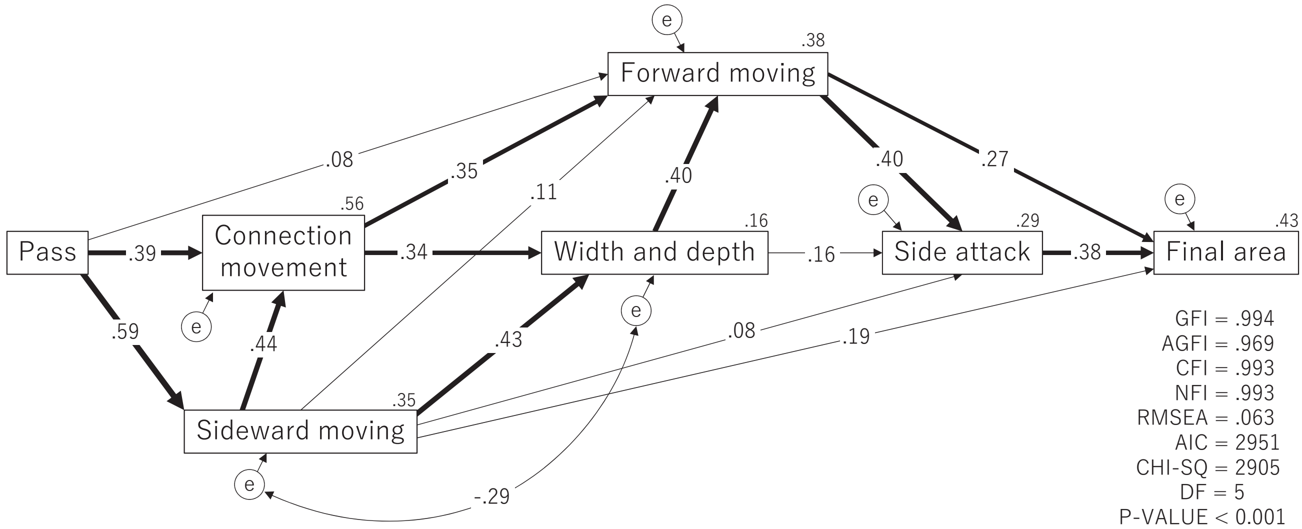
In the FS model of the play data (model FS-P), shown in **Figure 4**, the path coefficients were overall low to moderate. The paths with a moderate effect were “Pass → Connection movement” at 0.54, “Pass → Forward moving” at 0.40, “Connection movement → Sideward moving” at 0.52, “Forward moving → Side attack” at 0.40, and “Side attack → Final area” at 0.38.

4. Discussion

4.1. Achievement standard

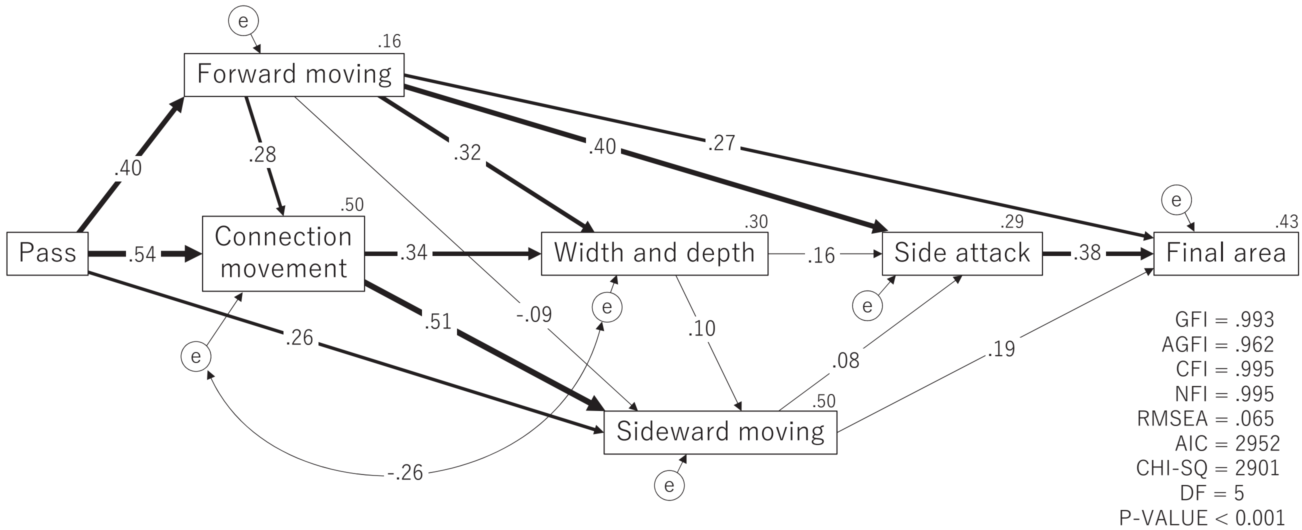
Soccer game performance data are ball touch and tracking data that are measured with quantitative scales, such as the distance covered and the number of passes. In the previous sections, the achievement standards to convert this type of data into binary data were calculated by classification binary tree analysis. These achievement standards are the strongest indicators of whether the player shoots the ball or not, and since the shooting rate was higher when most of the items were achieved, they made it possible to identify the level of attacking skills with the two values of “achieved” or “unachieved.”

Another advantage of achievement binary data is that it reduces the measurement cost. For example,



All path coefficients were significant (p < 0.001).

Figure 1 Result of the model SF-A (Sideward to forward with achievement data)



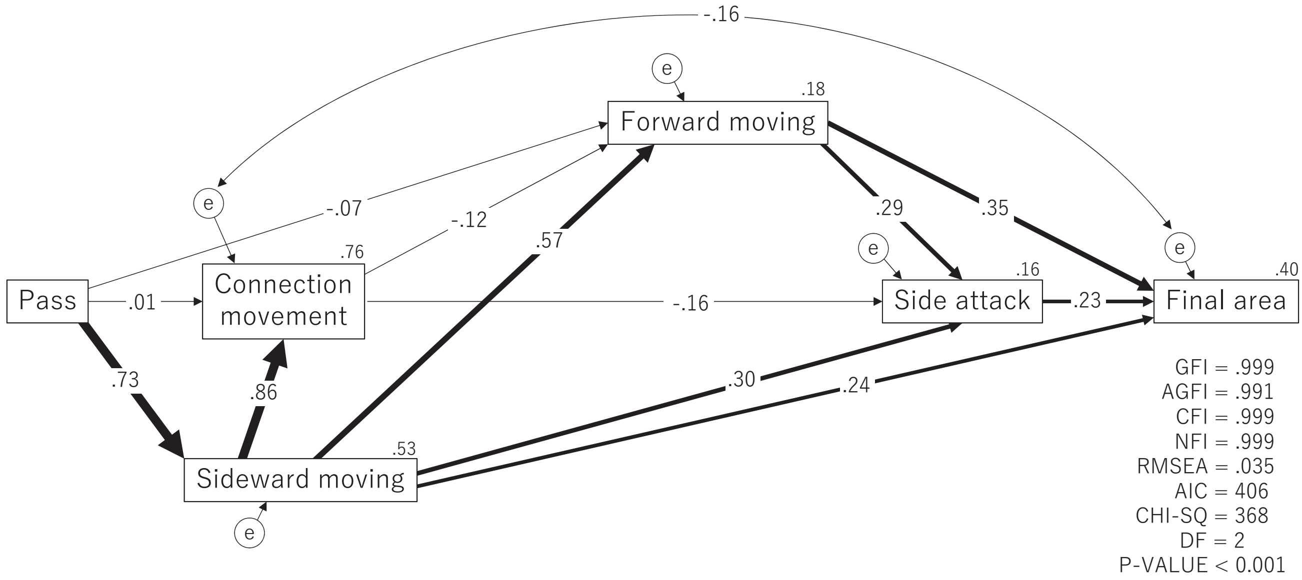
All path coefficients were significant (p < 0.001).

Figure 2 Result of the model FS-A (forward to sideward with achievement data)

since the achievement standard of "Pass" is four or more passes, when the number of passes in a single play reaches four, it is no longer necessary to look at the pass motions. The same applies to distance-related items; after the achievement standard is exceeded, there is no need to measure the distance covered anymore. Also, the use of a single unit makes it more practical for instruction fields.

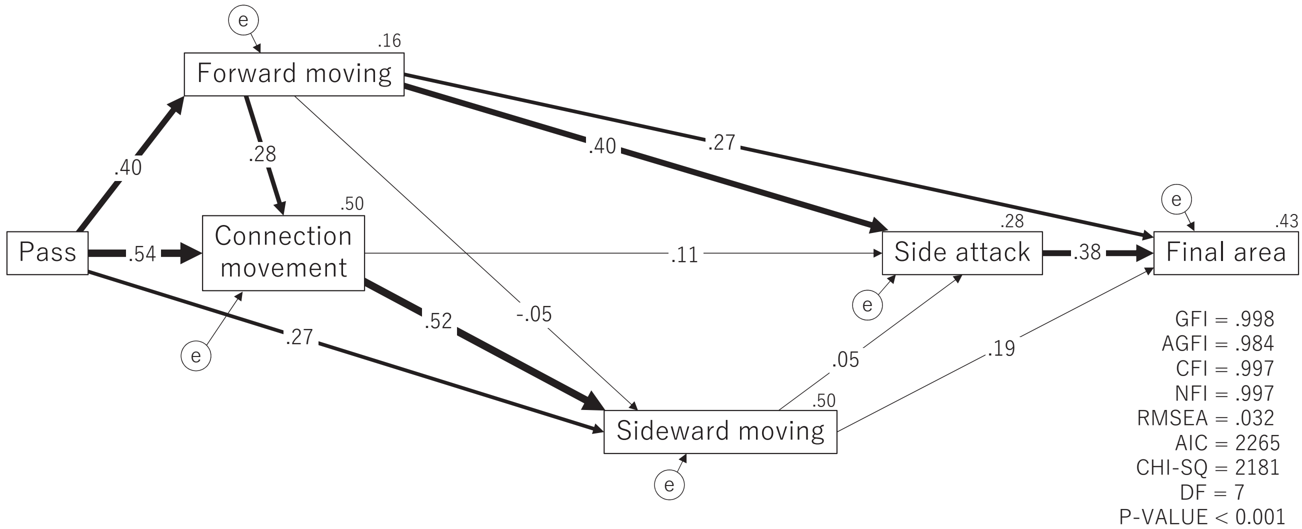
However, because the conversion of quantitative data into achievement binary data changes the characteristics of the data, it was necessary to verify whether the binary data truly reflect soccer skills. One problem that could happen in practice is, for example, that the values of opposite play styles like possession and counterattack are not registered properly. Possession-focused plays tend to have more passes and a longer attack time, while counterattack-minded plays tend to be the opposite. Consequently, when the variables "number of passes" and "attack time" are converted into binary — achieved or unachieved — it can only evaluate one of the two. Hence, if we can prove that it is valuable to use binary data, despite this kind of "binary conversion of opposite data problem," it may represent significant progress in measuring and evaluating game performance.

The objective of the attack in soccer is to score



All path coefficients were significant ($p < 0.001$).

Figure 3 Result of the model SF-P (Sideward to forward with play data)



All path coefficients were significant ($p < 0.001$).

Figure 4 Result of the model FS-P (forward to sideward with play data)

a goal, and since shooting is the only method to do so, it was appropriate to assign "Shoot" as a dependent variable when calculating the achievement standards. J.League's average shooting rate in 2011 was 8.4%, which means that, if the shooting rate when each measurement item is achieved exceeds 8.4%, the chances of a shot increases if that item is achieved. The item with the highest shooting rate when achieved was "Success of cross," at 70.5%. However, the number of times this item was achieved was small, at 3,442 plays (2.3%). Also, the shooting rate when "Success of through-ball" was achieved is relatively high at 35.1%, but the number of successful through-balls was small, at 6,534 (4.4%).

If the shooting rate is high when a measurement item is achieved, it is a strong indicator of whether there was a shot or not, which makes that item valuable. However, if the evaluation is based on items that are rarely achieved, it is necessary to measure a high volume of data, which ends up increasing the workload of the instruction field. Meanwhile, the shooting rate when "Penalty area" and "Vital area" were achieved were respectively 56.3% and 40.0%, and they were achieved a relatively high number of times, at 11,042 (7.5%) and 26,887 (18.3%). Therefore, these can be considered very good measurement items. However, in soccer, when a player penetrates the “Final area,” it is natural that the shooting rate increases. Therefore, the effectiveness as a measurement item group is only guaranteed if the shooting rate is high when the other items are achieved as well. To verify this aspect, we excluded “Penalty area” and “Vital area,” and 19 items had a shooting rate higher than 15%, and 38 items had a shooting rate higher than 10% when achieved. This revealed that the chances of a shot increase when the achievement standard of most of the measurement items is met.

4.2 Exploratory factor analysis

The exploratory factor analysis of the achievement data generated a simple structure with seven factors, with a factor loading of 0.48 as a standard. On the other hand, with the play data, six factors were found with a factor loading of 0.52 as a standard, but it was not a simple structure. Using the function PAHFA, it was possible to test all patterns that can be used in the factor extraction methods, rotation methods, and the number of factors. The reason the factor extraction methods were limited to two options — maximum likelihood method and weighted least squares — is that, with this, we could use the achievement data with a statistically correct procedure in the analysis of causal relationship after the exploratory factor analysis.

As for the selection of variables, even though it still can be improved, it is not realistic to test all possible variable combinations, since there are over 130,000. Therefore, in this study, two specialists (a soccer coach and a university faculty member specializing in data analysis) had extensive discussions and repeated the analysis multiple times to select the variables. Ultimately, they concluded that it would be difficult to find a more appropriate and simpler structure and adopted the six factors with a non-simple structure.

The cumulative contribution rate of the seven factors of the achievement data was 91%, which is a considerably high explanation rate. Meanwhile, the play data with six factors had a rate of 81%. The achievement data also showed a higher interfactorial correlation value. These results suggest that the conversion of quantitative variables into binary simplified complex information and clarified their correlation. In other words, the structure in which the achievement of a certain item is related to the achievement of another item is more clearly reflected in the achievement data than in the play data. Jo (2020) points out that a classification binary tree analysis by Gini impurity is likely to produce statistical differences in physical exercise capacity in the two groups after the classification, and the achievement data of this study followed the same trend. This result strongly supports the significance of measuring performance with binary data in a soccer skills evaluation.

4.3. Confirmatory factor analysis

The goodness of fit indices GFI, AGFI, CFI, and NFI are considered good when they are equal to or higher than 0.9. The goodness of fit is considered good if the RMSEA is less than 0.05, unfit if equal to or higher than 0.1, and intermediate if equal to or higher than 0.05 and less than 0.1. At the stage before the model modification based on modification indices, both the achievement and play data had a GFI, CFI, and NFI higher than 0.9. Also, the RMSEA was less than 0.1 in the achievement data but 0.1 or higher in the play data. The factor loading ranged from 0.53 to 0.98 in the achievement data, and from 0.46 to 1.08 in the play data. Judging that the RMSEA value was less than 1.0 and the factor loading did not contain unfit solutions, the achievement data model was the better model before the modification.

When the goodness of fit was maximized, the GFI, AGFI, CFI, and NFI of both the achievement and play data exceeded 0.999. In addition, the RMSEA was less than 0.01. This shows that, if the model is modified without considering its consistency, it is possible to increase the goodness of fit without a limit. However, previous studies indicate that those kinds of modifications are inadequate (Toyoda, 2012) and that studies containing arbitrary analyses that do not indicate the modification process or rationale are not highly regarded (Yoshida et al., 2020). Even in sports science, there have been studies that do not mention the modification rationale or do not show the modification path in the resulting diagram. Although the model of this study before the modification had the lowest level of goodness of fit and a non-ideal AGFI, it is acceptable goodness of fit for a model before the modification. Presumably, by following the correct model modification procedure, it is possible to increase the goodness of fit even further.

4.4. Causal structure of attacking skills subfactors

Since it was not possible to guarantee a normal distribution in the achievement data — because it is qualitative — it was necessary to apply a path analysis between factor scores by the maximum likelihood method in the causal structure analysis (Mitsunaga et al., 2005). However, since the variable “Shoot” is not treated as a factor in the exploratory factor analysis, it could not be inserted in the variance-covariance matrix later. For this reason, we analyzed which factor affects the variable “Shoot” the most, by logistic regression, and performed a path analysis with that factor as the final endogenous variable. This two-stage analysis was the result of prioritizing statistically correct procedures.

The logistic regression analysis, with “Shoot” as a dependent variable and the factor score of the achievement data as an independent variable, revealed that “Final area” had the highest impact. This result showed that the factor score of achievement data reflects a playing characteristic of soccer that the players can only have a shot if they get close to the goal. Nagelkerke R^2 is an indicator of the goodness of fit of models corresponding to the coefficient of determination in linear regression analysis, and it is adjusted so that the maximum value of Cox-Snell R^2 is 1. Since Nagelkerke R^2 does not have a typical standard that determines whether it is good or bad, it must be judged based on the characteristics of the data and model. The data used in this study is about offensive plays in soccer, which is a complex phenomenon, and there are likely to be more factors related to the attacking performance other than those seven. Considering this, an intermediary Nagelkerke R^2 value of 0.446 can be considered acceptable goodness of fit.

Since it became clear that penetration into the “Final area” was the factor with the strongest effect on whether the player would take a shot or not, the next step was to clarify what a player needs to do to penetrate the final area. In this study, there were two candidates for an alternative to SEM using the achievement data: the path analysis between factor scores using the maximum likelihood method (Mitsunaga et al., 2005) and the asymptotically distribution-free weighted least squares (Finney and DiStefano, 2006), but there is another method that can be applied to non-normality data called diagonally weighted least squares (DWLS).

However, the fa function in the psych package in R used in the function PAHFA cannot specify DWLS. Therefore, if DWLS was used in an analysis of the causal relationship, it would mean using a different estimation method than exploratory factor analysis, so this method was discarded. Ultimately, because the maximum likelihood method showed the most appropriate factor structure in the exploratory factor analysis, it was decided to use the path analysis using factor scores by the maximum likelihood method to identify the causal relationship between factors.

In soccer, a common attacking style is to pass the ball from one side to the other, close to the halfway line or even behind it, to shake the opponent team and create chances, and only then go forward. After entering the enemy territory, teams often turn the ball from one side to the other to create a gap. This was also mentioned by a former Japan national team player based on his experience. On the other hand, the style of stealing the ball and immediately attacking vertically is also possible in some situations, but this only occurs under specific conditions like counterattacks. For this reason, the sequence of “horizontal movement, then vertical movement” is the main attacking style in soccer. Therefore, we applied these attacking styles to the path diagram and created the “SF model,” as well as the “FS model,” which is less frequent, for comparison.

In the SF model of achievement data (model SF-A), a modification path was drawn between the error variables of “Sideward moving” and “Width and depth.” Since the factors that influence the horizontal distance covered can also act as factors that widen the attack, this modification path is appropriate in terms of content. In the FS model of achievement data (model FS-A), a modification path was drawn between the error variables of “Connection movement” and “Width and depth.” If movements that change the direction or movements to advance in the same direction are seen, they may increase the area of attack, so this modification path is also valid in terms of content. In the SF model of play data (model SF-P), a modification path was drawn between error variables of “Connection movement” and “Final area.” To penetrate the penalty area and vital area, it is necessary to break the opponent’s defense using movements that change the direction or advance in the same direction, so this modification path also makes sense in terms of content. Other modification candidates were also calculated, but since no appropriate reason was found for them, only one modification was made on each of the three models.

The results of the SF model of achievement data (model SF-A) indicated an attacking style that starts from a pass, followed by horizontal movement, connection movements, and wide attack, and ultimately, the team penetrates the final area using a side attack. The path coefficient from “Forward moving” to “Final area” was 0.27, indicating a slightly weak effect. This means that plays designed to penetrate the final area through vertical movements, such as through pass and dribbling, are performed a certain number of times. However, the path coefficient was not very high because dribbling and through passes do not occur very often.

Meanwhile, the results of the FS model of achievement data (model FS-A) revealed that horizontal movement is not a factor that directly affects side attack or penetration into the final area. For defenders, it is important to not let the opponents get close to the goal, so except when close to the goal, they must prioritize preventing vertical movements over horizontal movements. Therefore, the attacking side must pass the ball from one side to the other to create a gap, and only then attack vertically. The path analysis indicated that this characteristic was expressed numerically, which suggests that the causal structure of the achievement data is appropriate.

We performed a similar analysis on the play data to compare its causal structure with that of the achievement data. The SF model of play data (model SF-P) showed a strong causal relationship in the shift of play from “Pass” → “Sideward moving” → “Connection moving.” Also, the shift from “Sideward moving” to “Side attack” showed an intermediate causal relationship. Presumably, the reason these causal relationships were stronger than those of the SF model of achievement data (model SF-A) is that the play data could measure plays that did not meet the achievement standards. For example, the achievement standard of “Pass” is four times or more. However, in an actual match, it is possible to move horizontally with three or fewer passes. There are also likely to be cases of side attacks that are set up even though items related to horizontal movement do not reach their achievement standards. This capacity of measuring even detailed performances that do not meet the achievement standards is a property that differentiates play data from achievement data.

Meanwhile, the fact that no factor of the play data was affected by “Connection movement” and that there was a slightly strong effect from “Sideward moving” on “Forward moving” is a very different result from that of the SF model of achievement data (model SF-A). This means that the SF model of play data (model SF-P) could not correctly measure the fact that it is important that attacking plays contain movements that change the ball’s moving direction or expand in the same direction.

The path coefficient of the FS model of play data (model FS-P) does not differ from that of the FS model of achievement data (model FS-A). Hence, even though the SF model in play data (model SF-P) showed some advantages of quantitative data, it could not reflect “Connection moving” in the path diagram, which is an important element in the attack. Therefore, we concluded that the SF model of the achievement data (model SF-A) is a more appropriate causal structure.

However, when using achievement data, it is necessary to carefully discuss the “binary conversion of opposite data problem” mentioned above. In soccer, some studies point out that shorter attack time results in higher goal rates (Acar et al., 2009), while others state that a higher number of passes results in higher goal rates (Hughes and Franks, 2005), but this is because the difference between possession game and counterattack.

Since counterattacks cannot occur at positions far from the goal, stealing the ball at the attacking third (the area closer to the goal, when the pitch is divided in three) is a necessary condition, but the number of plays that fulfill that condition was small, at 18,744 (12.7%). Moreover, it is influenced by the number of players in the defense and the match’s flow, so the actual number of counterattacks may be much smaller. Also, the play data of this study has a high correlation between attack time and the number of passes, and the achievement standard is four or more passes for “Pass” and 10 seconds or longer for “Duration of attack.” Based on these, it is presumed that the achievement standards of this study, along with the results of factor analysis and path analysis, are based on possession play.

Rather than an average value, an achievement standard is a value divided to maximize the shooting rate, and it can only be obtained accurately with a relatively high frequency after branching. But since this study used big data of game performance in J.League, this issue was solved. By confirming the validity of achievement standards and achievement binary data in this study, it was possible to apply IRT to game performance data in offensive soccer plays.

5. Limitations of this study

The aspects that limit the generalization of the conclusion of this study include limitations related to the samples, measurement items, and the dataset set. This study focused on J.League in the year 2011, and if the same analysis methods were applied to more recent matches or leagues other than J.League, the results would not necessarily be valid. Also, the play data, which is the basis of the achievement data, is only ball touch data; that is, it does not contain any position information other than the player's position with the ball, which is collected by tracking. For this reason, it does not refer to spatial movements. Furthermore, since the achievement data is binary, it judges whether a measured value is large (high) or small (low). Two sets of opposing data — possession game and counterattack — were analyzed, and the one with a higher incidence rate was prioritized in the results.

6. Conclusion

A few years ago, soccer game performance started being measured as big data, and with it came the need to develop models, analyses, and management methods using that big data and return them to the fields. If game performance data of quantitative scale is converted into achievement binary, it not only reduces the measurement cost but also makes it possible to perform a criterion-referenced evaluation with IRT applied. However, since the conversion into qualitative data decreases the amount of information, it is necessary to re-confirm if it still reflects the way soccer skills are performed.

Therefore, this study sought to identify the causal structure of soccer attacking skills from achievement binary performance big data. To this end, we looked at large-scale game performance data of offensive plays in J.League and analyzed the achievement standard of each item to build achievement binary data. The conclusions obtained are as follows:

(1) The achievement standards of 44 measurement items were determined using J.League's big data about attacking performance in soccer.

(2) The achievement binary data has seven attacking skills subfactors involved in the attacking performance in soccer games, as well as causal relationships that indicate the attacking style, and the factor structure and causal structure were valid and compatible.

These findings revealed that attacking game performance data measured by achievement binary is valid data that reflects soccer attacking skills.

7. Future tasks

A future task is to apply IRT to achievement binary data and thereby, create criterion-referenced evaluation criteria for soccer attacking skills.

8. Acknowledgements

This research was funded by the Shizuoka Sangyo University special research support grant, with data provided by Data Stadium Inc. I would like to express my gratitude to both institutions.

References

Abe, K., Sakumura, T., and Kamakura, T. (2017). Proposal of skill evaluation index of baseball player using item response theory. Proceedings of the Institute of Statistical Mathematics, 65(2): 235-249. (in Japanese).

Acar, F., Yapicioglu, B., Arikan, N., Yalcin, S., Ates, N., and Ergun, M. (2009). Analysis of goals scored in the 2006 World Cup. In Reilly, T., and Korkusuz, F. (eds.) Science and Football VI (pp.235-242). London: Routledge.

Ando, K., Mishio, S., and Nishijima, T. (2018a). Analysis of items and characteristics of a criterion-referenced self-administered test utilizing video images for the development of a computerized adaptive test for tactical skills in soccer. Football Science, 15: 26-37.

Ando, K., Mishio, S., and Nishijima, T. (2018b). Validity and reliability of computerized adaptive test of soccer tactical skill. Football Science, 15: 38-51.

Aoyagi, R. (2002). Application of item response theory to subjective assessment of run performance in children. Japan Journal of Test and Measurement in Health and Physical Education, 2: 1-9. (in Japanese).

Aoyagi, R. (2005). Measurement of motor ability with item response theory. Fukuoka: Tokashobo. (in Japanese).

Aoyagi, R. (2006). A subjective evaluation of the throwing form of young children by item response theory. Japan Journal of Test and Measurement in Health and Physical Education, 6: 1-9. (in Japanese).

Breiman, L., Friedman, J. H., Olsehn, R. A., and Stone, C. J. (1984). Classification and regression trees. UK: Chapman and Hall/CRC.

Dmonte, R., and Dmello, A. (2017). Big data in sports: Leverage big data in sports: An insight using SAP HANA. IJERT, 6: 380-383.

Finney, S. J., and DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock and R. O. Mueller (Hrsg.). Structural equation modeling: A second course (pp.269–314). Greenwich, Connecticut: Information Age Publishing.

Hughes, D., and Franks, I. (2005). Analysis of passing sequences shots and goals in soccer. J. Sports Sci., 23(5): 509-514.

Jo, H., Yokoo, T., Ando, K., Nishijima, T., Kumagai, S., Naomoto, H., Suzuki, K., Yamada, H., Nakano, T., and Saito, K. (2014). Attack power index of players and teams in J.League. Research on Sports Data Analytics: Theory, Methodology, and Applications, 1:21-26. (in Japanese).

Jo, H. (2020). How to use decision tree analysis to determine division points. Children, growth and development, 18:211-216. (in Japanese).

Jo, H., Matsuoka, H., Ando, K. and Nishijima, T. (2022a). Construction of offensive play measurement items and shot prediction model applying machine learning in Japan professional football league, Football Science, 19:1-21.

Jo, H., Nakanishi, K., and Aoki, M. (2022b). Factor structure of game performance big data in offensive play of Japan professional football league: Applying a practical program in exploratory factor analysis. Sports and Human in Bulletin of Shizuoka Sangyo University, 6. (in press). (in Japanese).

Matsuo, T., and Nakamura, T. (2002). Factor analysis nobody taught (p.75). Kyoto: Kitaouji publishing. (in Japanese).

Matsuoka, H., Tahara, Y., Ando, K., and Nishijima T. (2020). Development of criterion-referenced measurement items for soccer defensive tactical play from tracking data. Football Science, 17:29-40.

Memmert, D., and Raabe, D. (2019). Revolution im profifußball: Mit big data zur spielanalyse 4.0. (p.5). Berlin: Springer.

Mitsunaga, H., Hoshino, T., Shigemasu, K., and Mayekawa, S. (2005). Parameter estimation of structural equation models using latent variables scores. The Japanese Journal of Behaviormetrics, 32:21-33. (in Japanese).

Nicholas, W., Stephen, S., and Joris, Drayer. (2021). Big data and analytics in sport management. Journal of Sport Management, 35: 197-202.

Oe, J., Usui, S., Okihara, K., Shiokawa, M., Kan, A., Kajiyama, T., and Kurokawa, T. (2007). Quantitative analysis of the attacking performance in soccer games. Journal of Methodology of Ports, 20: 1-14.

Ohtomo, K. (1996). Introduction to item response theory: New analysis method of language test data. Tokyo: Taishukan publishing. (in Japanese).

Ono, M., Jo, H., Ohyama, K., and Nishijima, T. (2014). Achievement evaluation standard for men's discus throw. Japan Journal of Test and Measurement in Health and Physical Education, 14: 1-10. (in Japanese).

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (accessed 2021-10-19).

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish institute for educational research.

Revelle, W. (2020). psych: Procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3. (accessed 2021-11-4).

Suzuki, K., Yamada, Y., Osako, T., Takahashi, S., and Nishijima, T. (2000). Measurement of the shoot skill in forward players from game performance. Medicine and Science in Soccer, 20:37-41. (in Japanese).

Suzuki, K., and Nishijima, T. (2002). Causal structure of attacking skill in soccer games. Jpn J. Phys. Educ. Health Sport Sci., 47(6): 547-567. (in Japanese).

Toyoda, H. (2012). Section 9: Confirmatory factor analysis. In Introduction to factor analysis: Latest data analysis learn with R (pp.212-240). Tokyo: Tokyo-tosho publishing. (in Japanese).

Viktor, S., and Kenneth, C. (2013). Letting the data speak. In Big data: A revolution that will transform how we live, work, and think. (pp.6-12). London: John Murra.

Yamada, H., Suzuki, K., Osako, T., Takahashi, S., and Nishijima, T. (2000). Measurement of defensive pressure from game performance. Medicine and Science in Soccer, 20:32-36. (in Japanese).

Yoshida, T., Murai, J., Usami, S., Shojima, K., Oshio, A., Suzuki, M., and Shiina, K. (2020). What has structural equation modeling brought to psychology? The Annual Report of Educational Psychology in Japan, 59:292-303. (in Japanese).

Yves, R. (2012). Lavaan: An R package for structural equation modeling. J. Stat. Softw., 48(2):1-36.

Name:
Hirotaka Jo

Affiliation:
Department of Sports Science, Shizuoka Sangyo University and Doctoral Program in Physical Education Health and Sport Sciences, University of Tsukuba

Address:
1572-1 Oowara, Iwata-city, Shizuoka-pref 438-0043 Japan

Brief Biographical History:

- 2014- Health and physical education teacher, Junior & Senior High School at Komaba, University of Tsukuba.
- 2019- Associate lecturer, Shizuoka Sangyo University.
- 2022- Associate professor, Shizuoka Sangyo University.

Main Works:

- Jo, H., Oosawa, K., Mishio, S., Ando, K., Suzuki, K., and Nishijima, T. (2017). Development of optimization algorithm for attack play in football., Proceedings of the Institute of Statistical Mathematics, 64: 309-321. (in Japanese).
- Jo, H., Matsuoka, H., Ando, K. and Nishijima, T. (2022). Construction of offensive play measurement items and shot prediction model applying machine learning in Japan professional football league, Football Science, 19:1-21.

Membership in Learned Societies:

- Japanese Society of Science and Football
- Japan Society of Physical education, Health and Sport Sciences
- Japanese Society of Test and Measurement in Health and Physical Education