Why is soccer so popular:
Understanding underdog achievement and randomness in team ball sports

L. N. Vicente[1]   T. N. Alleck[1]   T. Giovannelli[2]   R. Mitchell[1]   O. Remen[1]

February 6, 2025

Abstract

In this paper, we examine team ball sports to investigate how the likelihood of weaker teams winning against stronger ones, referred to as underdog achievement, is influenced by inherent randomness factors that affect match outcomes in such sports. To address our research question, we collected data on match scores and computed corresponding team rankings from major international competitions (World cups or Olympic games) for 12 popular team ball sports: basketball, cricket, field hockey, futsal, handball, ice hockey, lacrosse, roller hockey, rugby, soccer, volleyball, and water polo. Then, we developed an underdog achievement score to identify the sports with the highest occurrences of weaker teams prevailing over stronger ones, and we designed a randomness model consisting of factors that contribute to unexpected match outcomes within each sport. Our findings indicate that soccer is among the sports in which a weaker team is most likely to win. Through principal component analysis (PCA) and correlation analysis, we demonstrate that our randomness model can explain such a phenomenon, showing that the underdog achievement can be attributed to numerous factors that can randomly influence match outcomes.

1 Introduction

Team ball sports have been the subject of growing research over the past decade, and most of the papers are within the domain of sports medicine literature [15]. In our work, we do not focus on medicine applications. Instead, we aim to provide novel insights into team ball sports by determining which sports are more likely to see weaker teams win, referred to as underdog achievement, and which randomness factors contribute to such an outcome. Since certain teams consistently outperform others in all sports, it is clear that outcomes of matches are not purely random. The question of whether a team wins due to chance or their own skills has been discussed in non-scholarly books. For example, in [14], the authors claim that soccer, which is universally recognized as the world's most popular sport [3], is also the most random, and its inherent randomness is what makes soccer so popular. Similar conclusions were confirmed in the academic literature by [2], where the authors analyzed the English Football Association and four major North American professional sports leagues (MLB for baseball, NBA for basketball,

---

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582
[2]Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45221-0072
Emails: lnv@lehigh.edu, tna324@lehigh.edu, giovanto@ucmail.uc.edu, rgm424@lehigh.edu, orr224@lehigh.edu

NFL for American football, and NHL for hockey), finding that soccer is the sport with the most random outcomes. In general, the randomness element adds excitement and unpredictability, making a sport enjoyable to watch. However, one should also notice that excessive randomness can diminish interest, as viewers prefer a balance between unpredictability and skill [13].

A novel research question is whether a weak team is more likely to win than a strong one as a consequence of certain random and situational conditions. Such a question has recently emerged as a research focus and has been studied by [16], where the authors use data from the English Premier League to show that the influence of randomness on goals in soccer decreases as the match progresses. Such a decreasing trend was observed to be disadvantageous for weaker teams, as they rely more on randomness to score. The study in [16] also identifies variables of randomness that affect the outcome during a match and cannot be entirely attributed to skills. Examples of such variables include the degree of involvement of the defending team and the chances to score goals from outside the penalty area. Additionally, the analysis includes situational variables that may influence the outcome by affecting the motivation of players, such as match location and current score. We conclude our literature review with [12], where a Bayesian state-space model was proposed to study the randomness in match outcomes for four major North American professional sports leagues (once again, MLB, NBA, NFL, and NHL). Probability-based metrics derived from betting market data were used to quantify the influence of chance on the outcomes. Their findings indicate that the MLB and NHL exhibit the highest levels of randomness in match outcomes (and we found a similar result for ice hockey in our study). However, soccer was not included in their analysis.

Previous studies examining underdog achievement have focused on a limited range of team ball sports and have not systematically identified the randomness factors contributing to underdog victories. The ultimate goal of our paper is to investigate how the likelihood of weaker teams winning against stronger ones is influenced by inherent randomness factors in team ball sports. To achieve our goal, we selected major international competitions for 12 popular team ball sports: basketball, cricket, field hockey, futsal, handball, ice hockey*, lacrosse, roller hockey, rugby, soccer, volleyball, and water polo. The official names of the competitions selected for each sport are included in Table 1. Note that such competitions are all men's events, for which more data is available. We notice that certain popular team ball sports, such as American football, baseball, and tennis, were not included in our analysis. This decision was made due to either the absence of international competitions for these sports or the smaller size of their teams compared to the sports considered in this paper (for instance, in tennis, teams consist of at most two players, whereas all the other sports discussed in our paper involve teams with many more than two players).

Our main contributions can be summarized as follows:

- We collected match score data from major international competitions (World cups or Olympic games) held between 1970 and 2023, and we computed corresponding team rankings for each edition of the competitions in Table 1. This data represents valuable information for researchers in the field of sports analytics.

- We developed an underdog achievement score to determine the sports with the highest and lowest occurrences of weaker teams defeating stronger ones when focusing on a much broader range of team ball sports than the ones considered in the literature. In accordance

_____

*We include ice hockey among the team ball sports, even though it technically uses a puck.

with the limited existing literature (again, see [2, 14]), soccer is among the sports with the highest underdog achievement.

- We designed a randomness model consisting of 13 factors that contribute to unexpected match outcomes within each sport, providing quantitative values for each of the factors.

- We performed principal component analysis (PCA) and correlation analysis to identify the randomness factors with the greatest impact on underdog achievement and demonstrate that our randomness model can explain underdog achievement.

Our paper is organized as follows. In Section 2, we detail our data collection process. In Section 3, we develop an underdog achievement score and in Section 4, we present our randomness model. In Section 5, we perform PCA and correlation analysis to demonstrate how our randomness model can explain underdog achievement. In Section 6, we conclude our paper with some remarks and ideas for future work.

Sport	Major International Competition
Basketball	Summer Olympic Games
Cricket	ICC Men's Cricket World Cup
Field Hockey	Men's FIH Hockey World Cup
Futsal	FIFA Futsal World Cup
Handball	Summer Olympic Games
Ice Hockey	Winter Olympic Games
Lacrosse	World Lacrosse Men's World Cup
Roller Hockey	World Skate Roller Hockey World Cup
Rugby	Rugby World Cup
Soccer	FIFA World Cup
Volleyball	FIVB Volleyball Men's World Cup
Water Polo	FINA Men's Water Polo World Cup

Table 1: Major international competitions selected for the team ball sports included in our paper.

2 Data collection: Match scores and team rankings

To perform our analysis, for each sport, we collected real data on match scores and computed corresponding team rankings for the major international sports competitions in Table 1. We selected such competitions due to the ease of access and availability of their data. The complete table with the years of the editions of each competition is provided in Table 6 of Appendix A. We have arbitrarily included editions from 1970 onward to avoid introducing bias into our analysis, as earlier editions may exhibit different patterns compared to more recent ones. To obtain match score data for each edition, we conducted web scraping of match information from Wikipedia pages. All this data was aggregated into a match score dataset, which contains information related to individual matches, including the names of the two opposing teams and their respective scores. Given the match score dataset, we then computed a team ranking for each edition. Finally, for each competition, we aggregated the team rankings across all the

$\mathcal{S}$	Set of sports.
$\mathcal{E}$	Set of editions for the given competition.
$\mathcal{P}_e = \{p_1, p_2, \ldots, p_{N_e}\}$	Set of all teams playing in edition $e \in \mathcal{E}$.
$\mathcal{M}_e \subseteq \mathcal{P}_e \times \mathcal{P}_e$	Set of matches in edition $e \in \mathcal{E}$.
$\mathrm{sc}_{ij}^e(i)$, $\mathrm{sc}_{ij}^e(j)$	Scores obtained by teams i and j when facing each other, where $(i,j) \in \mathcal{M}_e$.
$\mathcal{D}_e$	Match score dataset for edition $e \in \mathcal{E}$, as defined in (2.1).
$\mathcal{R}_e$	Team ranking for edition $e \in \mathcal{E}$, as defined in (2.2).
$\mathcal{E} = (e_1, e_2, \ldots, e_{|\mathcal{E}|})$	Set of editions written as an ordered list of elements, where e_1 is the earliest edition and $e_{|\mathcal{E}|}$ is the most recent edition.
$\mathcal{P}_{\leqslant e}$	Union of the sets of teams up to edition $e \in \mathcal{E}$, as defined in (2.3).
$N_{\leqslant e}$	Cardinality of the set $\mathcal{P}_{\leqslant e}$.
$c(i, \mathcal{R}_e)$	Position of a team $i \in \mathcal{P}_e$ in the team ranking $\mathcal{R}_e$.
$\mathcal{W}_{\leqslant e}$	Weighted team ranking up to edition $e \in \mathcal{E}_s$, as defined in (2.5).
$wr_{\leqslant e}$	Sorting function used to assign each team to their position in $\mathcal{W}_{\leqslant e}$, as defined in (2.4).
$c(i, \mathcal{W}_{\leqslant e})$	Position of a team $i \in \mathcal{P}_{\leqslant e}$ in the weighted team ranking $\mathcal{W}_{\leqslant e}$.
$|c(i, \mathcal{R}_e) - c(j, \mathcal{R}_e)|$	Rank difference between teams $i \in \mathcal{P}_e$ and $j \in \mathcal{P}_e$ in $\mathcal{R}_e$.
$|c(i, \mathcal{W}_{\leqslant e}) - c(j, \mathcal{W}_{\leqslant e})|$	Rank difference between teams $i \in \mathcal{P}_{\leqslant e}$ and $j \in \mathcal{P}_{\leqslant e}$ in $\mathcal{W}_{\leqslant e}$.
τ	Rank difference threshold used to identify weak teams, as defined in (3.1).
λ	Decay factor used in $wr_{\leqslant e}$ to determine past edition relevance.

Table 2: Notation.

edition years included in Table 6 into a weighted team ranking. Our code is publicly available on GitHub.[†]

We will now introduce some general notation that will allow us to formally describe the match score dataset, the team ranking for each edition, and the weighted team ranking. Denoting as $\mathcal{S}$ the set of sports, let $\mathcal{E}_s$ be the set of editions for the competition selected for sport $s \in \mathcal{S}$ in Table 1 (one can think of such a set as a set of edition years) and let $\mathcal{P}_e^s = \{p_1, p_2, \ldots, p_{N_e^s}\}$ be the set of all teams playing in edition $e \in \mathcal{E}_s$, where N_e^s is the total number of teams in that edition. We will denote as $\mathcal{M}_e^s \subseteq \mathcal{P}_e^s \times \mathcal{P}_e^s$ the set of matches in edition e, represented as a set of tuples (i, j), where i and j are opposing teams belonging to $\mathcal{P}_e^s$. For the rest of this section, we will omit the subscript s to simplify the notation. All the notation used in our paper is summarized in Table 2.

Match score dataset. For any edition $e \in \mathcal{E}$ and match $(i, j) \in \mathcal{M}_e$, let $\mathrm{sc}_{ij}^e(i)$ denote the score that team i obtained when playing against team j (for example, if "5-6" is the outcome of the match (i, j), then $\mathrm{sc}_{ij}^e(i) = 5$ and $\mathrm{sc}_{ij}^e(j) = 6$). The match score dataset for edition e can be represented by the following set

$$\mathcal{D}_e = \{(e, i, j, \mathrm{sc}_{ij}^e(i), \mathrm{sc}_{ij}^e(j)) \mid (i, j) \in \mathcal{M}_e\}. \tag{2.1}$$

To populate such a dataset, we web-scraped information related to individual matches from the Wikipedia pages corresponding to each edition year of a competition. All sports competitions

---

[†]https://github.com/thaksheel/randomness-team-ball-sports.git

typically include a group stage, where teams are divided into groups and each team plays against the others in its group to collect the maximum number of points and advance in the competition, and a knockout stage (or bracket stage), where teams are eliminated from the competition if they lose a match. The knockout stage typically consists of the following additional phases: rounds of 16, quarterfinals, semifinals, and finals. For more details, we refer to [1].

Note that the match score dataset is a *convenience sample* [5, 7], meaning it consists of data for specific competitions that was conveniently accessible and ready to be collected, rather than a dataset chosen through a random or probabilistic selection process from the population of interest, which includes all sports matches. This can introduce certain limitations due to the potential biases in the types of matches included in the sample, which may not fully represent the entire population. Despite such limitations, convenience sampling is the most common sampling method due to the impracticality of accessing the entire population [4].

Team rankings. Based on the match score dataset $\mathcal{D}_e$ in (2.1), for each edition $e \in \mathcal{E}$, we generated a team ranking by sorting teams based on the following criteria, listed in order of priority: number of matches played, number of victories, number of draws, number of losses, and total score across all matches. The sorting was in descending order for each criterion except for the number of losses, for which an ascending order was used. The choice to use the number of matches played as a sorting criterion is due to the fact that teams reaching the final stages of a competition typically play the largest number of matches, reflecting their strength. In cases where two teams have played the same number of matches, the team with more victories is ranked higher. If the number of victories is equal, the team with more draws is ranked higher, and so forth. For each edition e, the team ranking obtained through the sorting criteria above is represented as an ordered list of teams

$$\mathcal{R}_e = (i_1, i_2, \ldots, i_{N_e}), \tag{2.2}$$

where $i_j \in \mathcal{P}_e$ for any $j \in \{1, 2, \ldots, N_e\}$.

An alternative to the team ranking logic adopted in our paper would be to use official team rankings from major international competitions. However, the availability of official ranking data is limited for many sports and competition editions. Official team rankings are mainly available for soccer and rugby, while rankings for other sports are often inconsistently documented or inaccessible. An advantage of using our team ranking logic is that it applies uniformly across all sports, despite the different rules and official ranking systems used in different competitions. We acknowledge that our team ranking logic does not account for the importance of specific matches within a competition, a factor typically considered in official rankings. In our approach, all losses and draws are weighted equally, regardless of the match's significance.

Weighted team ranking. The weighted team ranking aggregates team rankings from the earliest available edition up to a given edition. When referring to the weighted team ranking, we will explicitly write the set of editions as an ordered list of elements as follows $\mathcal{E} = (e_1, e_2, \ldots, e_{|\mathcal{E}|})$, where e_1 is the earliest available edition and $e_{|\mathcal{E}|}$ is the most recent edition. In such a case, for any $h \in \{1, 2, \ldots, |\mathcal{E}|\}$, we will denote as

$$\mathcal{P}_{\leqslant e_h} = \cup \{\mathcal{P}_{e_{\bar{h}}} \mid e_{\bar{h}} \in \mathcal{E} \text{ and } \bar{h} \leq h\} \tag{2.3}$$

the union of the sets of teams from edition e_1 to edition e_h. We will denote as $N_{\leqslant e_h}$ the cardinality of the set $\mathcal{P}_{\leqslant e_h}$.

To build the weighted team ranking that aggregates team rankings up to an arbitrary edition $e \in \mathcal{E}$, denoted as $\mathcal{W}_{\leqslant e}$, we use a sorting function $wr_{\leqslant e} : \mathcal{P}_{\leqslant e} \to \mathbb{R}$ to assign each team to their position in such a weighted team ranking. For any $i \in \mathcal{P}_{\leqslant e}$, the higher the value of $wr_{\leqslant e}(i)$, the higher the position of team i in the weighted team ranking. To show how we compute the weighted team ranking $\mathcal{W}_{\leqslant e}$, let us denote as $c(i, \mathcal{R}_e) \in [1, N_e]$ the position of a team $i \in \mathcal{P}_e$ in the team ranking $\mathcal{R}_e$ in (2.2). For any $e_h \in \mathcal{E} = (e_1, e_2, \ldots, e_{|\mathcal{E}|})$, with $h \in \{1, 2, \ldots, |\mathcal{E}|\}$, we compute $wr_{\leqslant e_h}(i)$ as follows

$$
wr_{\leqslant e_h}(i) \;=\; \begin{cases} (N_{e_h} - c(i, \mathcal{R}_{e_h}))/N_{e_h}, & \text{if } i \in \mathcal{P}_{e_h} \text{ and } h = 1, \\ (N_{e_h} - c(i, \mathcal{R}_{e_h}))/N_{e_h} + \lambda\, wr_{\leqslant e_{h-1}}(i), & \text{if } i \in \mathcal{P}_{e_h} \text{ and } h \in \{2, \ldots, |\mathcal{E}|\}, \\ 0, & \text{if } i \notin \mathcal{P}_{e_h}, \end{cases} \quad (2.4)
$$

where $\lambda \in [0, 1]$ is a decay factor that dictates the rate at which past editions become irrelevant. Note that the division by N_{e_h} in (2.4) ensures that the weighted team ranking is not affected by the number of teams. This is important to avoid introducing bias in the analysis, as the same competition may feature a different number of teams in different editions, and competitions across different sports may have significantly different numbers of teams.

The weighted team ranking up to edition e is represented as an ordered list of teams

$$
\mathcal{W}_{\leqslant e} = (i_1, i_2, \ldots, i_{N_{\leqslant e}}), \tag{2.5}
$$

where $i_j \in \mathcal{P}_{\leqslant e}$ for any $j \in \{1, 2, \ldots, N_{\leqslant e}\}$ and $wr_{\leqslant e}(i_j) \geq wr_{\leqslant e}(i_{j+1})$ for any $e \in \mathcal{E}$ and $j \in \{1, 2, \ldots, N_{\leqslant e} - 1\}$.

3 Underdog achievement

To quantify the underdog achievement, we first need to determine criteria that allow us to distinguish weak teams from strong ones. As noticed in the literature related to soccer [16], determining a team's strength is a difficult task because of the interaction between skills and randomness. Different approaches have been proposed to evaluate a team's strength, such as using the positions of teams in team rankings [6], the total number of points scored in a competition [8], ELO-ratings [9], or betting odds [16]. In this paper, weak teams are identified based on their positions in the weighted team ranking described in Section 2.

Identifying weak teams. Given the weighted team ranking in (2.5), one can consider two strategies to identify weak teams. A first strategy consists of considering the top $p\%$ teams in the weighted team ranking as strong and the bottom $p\%$ teams as weak. However, in our case, this strategy proved unsuccessful because teams in the bottom $p\%$ rarely, if ever, defeat teams in the top $p\%$, regardless of the sport. When $p = 50$, for some sports, it occasionally happens that teams in the bottom half defeat teams in the top half, while in others, it never happens. Nevertheless, this is just noise occurring due to teams in mid-ranking positions, and so it is not a reliable indication of weak teams prevailing over strong ones. Therefore, we considered a second strategy, described below.

The approach used in our paper consists of comparing the positions of two teams in the weighted team ranking and considering as weak the team that is ranked significantly lower, if it exists. In other words, there must be a relatively high difference in positions between the

teams in the weighted ranking to consider the lower-ranked team as weak. Using the notation introduced in Section 2, for each edition $e \in \mathcal{E}$, we denote as $c(i, \mathcal{W}_{\leqslant e}) \in [1, N_{\leqslant e}]$ the position of a team $i \in \mathcal{P}_{\leqslant e}$ in the weighted team ranking $\mathcal{W}_{\leqslant e}$. Given two teams i and j in $\mathcal{P}_{\leqslant e}$, we refer to $|c(i, \mathcal{W}_{\leqslant e}) - c(j, \mathcal{W}_{\leqslant e})|$ as the rank difference between teams i and j in the weighted team ranking $\mathcal{W}_{\leqslant e}$. Given a match $(i, j) \in \mathcal{M}_e$ between teams i and j in edition e, we identify i as a weak team based on $\mathcal{W}_{\leqslant e}$ if

$$c(i, \mathcal{W}_{\leqslant e}) \ \leq \ c(j, \mathcal{W}_{\leqslant e}) - \tau, \tag{3.1}$$

where τ is a positive threshold depending on the sport, to be determined. Note that (3.1) implies that the rank difference $|c(i, \mathcal{W}_{\leqslant e}) - c(j, \mathcal{W}_{\leqslant e})|$ is greater than τ. When (3.1) is not satisfied and, therefore, two teams have a rank difference less than or equal to the threshold τ, we assume that such teams have similar strengths and we do not classify either of them as weak.

Underdog achievement score.   Recall $\mathcal{E} = (e_1, e_2, \ldots, e_{|\mathcal{E}|})$, where e_1 is the earliest available edition and $e_{|\mathcal{E}|}$ is the most recent edition, and recall the definition of weak team based on $e \in \mathcal{E}$ in (3.1). For any $h \in \{2, \ldots, |\mathcal{E}|\}$, we define the underdog achievement score for edition $e_h \in \mathcal{E}$ as follows

$$\mathrm{UAS}_{e_h} \ = \ \frac{\text{Number of victories or draws in edition } e_h \text{ by weak teams based on } \mathcal{W}_{\leqslant e_{h-1}}}{\text{Number of matches in edition } e_h \text{ with a weak team based on } \mathcal{W}_{\leqslant e_{h-1}}}. \tag{3.2}$$

Note that in (3.2), weak teams are identified using a weighted team ranking that incorporates all past editions except the current one to avoid biasing the results. One can interpret (3.2) as the probability that a historically weaker team wins against a historically stronger one in a certain edition of a competition. For each sport, the average underdog achievement score across all editions is given by

$$\mathrm{UAS} \ = \ \frac{1}{|\mathcal{E}| - 1} \sum_{h=2}^{|\mathcal{E}|} \mathrm{UAS}_{e_h}. \tag{3.3}$$

Numerical results.   In this subsection, we first perform a rank difference analysis to determine the value of the threshold τ, used to identify weak teams in (3.1), which affects the computation of UAS_{e_h} and UAS in (3.2) and (3.3), respectively. Then, we quantify UAS for each sport. For each edition $e \in \mathcal{E}$, we recall that $c(i, \mathcal{R}_e) \in [1, N_e]$ denotes the position of a team $i \in \mathcal{P}_e$ in the team ranking $\mathcal{R}_e$. Given two teams i and j in $\mathcal{P}_e$, we refer to $|c(i, \mathcal{R}_e) - c(j, \mathcal{R}_e)|$ as the rank difference between teams i and j in the team ranking $\mathcal{R}_e$.

Figure 1 represents box plots showing the distribution of the rank differences between teams i and j in the team ranking $\mathcal{R}_e$ across all matches for all sports and editions, i.e., $|c(i, \mathcal{R}_e) - c(j, \mathcal{R}_e)|$ for all $(i, j) \in \mathcal{M}_e$, $e \in \mathcal{E}_s$, and $s \in \mathcal{S}$ (where $\mathcal{E}_s$ is the set of editions for sport s). The medians of the rank difference distributions across all sports range from 2 for water polo to 8 for soccer. In particular, for all sports except soccer, the median of the corresponding rank difference distribution is less than or equal to 5. For soccer, the rank differences range from 1 to 30, and approximately half of the soccer matches occurred with a rank difference less than or equal to 8. For the rest of the paper, we set the threshold τ in (3.1) equal to the median of the corresponding rank difference distribution for each sport (see Table 3). Note that the rank difference distribution is more influenced by the specific characteristics of the competition system (e.g., number of teams in the tournament, number of groups in the preliminary phase)
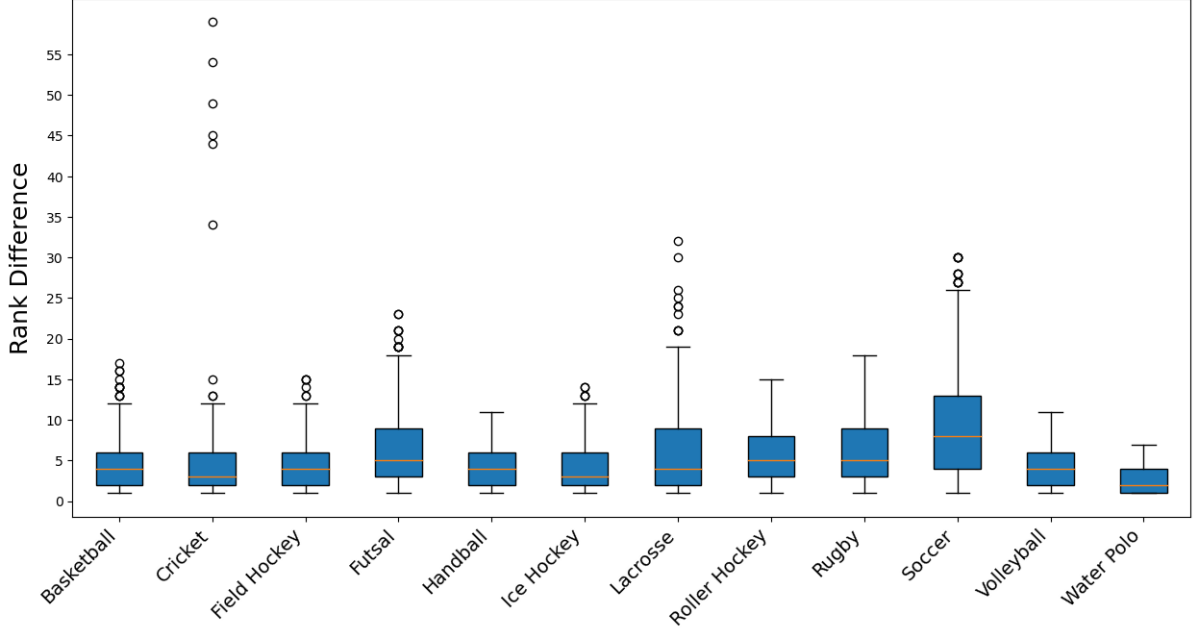
Figure 1: Box plot showing the distribution of rank differences between teams across all matches for each team ball sport included in our paper.

than by the sport itself. As the number of teams in a competition increases, the spread of rank differences tends to grow.

Figure 2 contains box plots representing the distribution of UAS_{e_h} values across all editions $e_h \in \mathcal{E}$ for each sport for three different values of the weight λ in (2.4), i.e., $\lambda \in \{0, 0.5, 1\}$. One can observe that regardless of the value of λ, basketball, field hockey, ice hockey, soccer, and water polo have consistently high values of UAS_{e_h} compared to the other sports, while lacrosse, roller hockey, and rugby have consistently low values of UAS_{e_h}. Water polo is the sport for which the UAS_{e_h} distribution has the highest median, followed by soccer. Conversely, lacrosse, roller hockey, and rugby are the sports for which the UAS_{e_h} distribution has the lowest median. Similar results can be observed in terms of the average underdog achievement score in (3.3) from Figure 3, which shows an upper triangular pairwise difference matrix representing the differences in UAS values for all pairs of sports. Note that in Figure 3, the weight λ in (2.4) is set to 0.5 (the results for $\lambda = 0$ and $\lambda = 1$ are included in Appendix B, as they lead to similar conclusions as $\lambda = 0.5$). The exact UAS values for each sport for the three different values of λ are reported in Table 3. As the value of λ decreases, indicating that past editions are weighted less in the weighted team ranking, the differences in UAS values between basketball on one side and lacrosse, roller hockey, and rugby on the other become less pronounced. This suggests that in basketball, there were strong teams in the past that have progressively weakened, which explains the high underdog achievement when past editions are excluded.

Figure 4 includes graphs that depict the evolution of UAS_{e_h} over time for each sport for the same three different values of the weight λ. For cricket, volleyball, and soccer, one can observe that in the early 1970s, the UAS_{e_h} values were low but began to increase over time. Interestingly, these years correspond to a period when the number of teams in the competitions was lower, as indicated in [1, Tables 5 and 8]. This suggests that with more teams in a competition, there
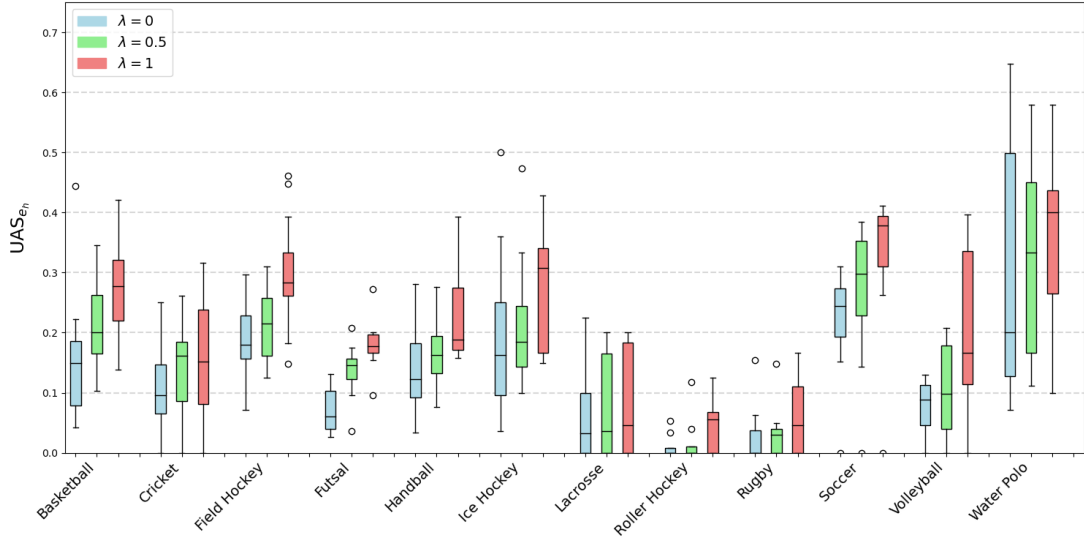
8

Figure 2: Box plot showing the UAS_{e_h} distribution for each sport for three different values of the weight λ in (2.4).

Sport	τ	UAS		
		$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$
Basketball	4	0.28	0.21	0.15
Cricket	3	0.16	0.13	0.10
Field Hockey	4	0.30	0.21	0.19
Futsal	5	0.18	0.14	0.07
Handball	4	0.23	0.16	0.14
Ice Hockey	3	0.28	0.21	0.20
Lacrosse	4	0.08	0.08	0.06
Roller Hockey	5	0.05	0.02	0.01
Rugby	5	0.06	0.04	0.03
Soccer	8	0.33	0.27	0.22
Volleyball	4	0.19	0.11	0.08
Water Polo	2	0.37	0.33	0.31

Table 3: Threshold τ and impact of λ on the UAS values for each team ball sport included in our paper.
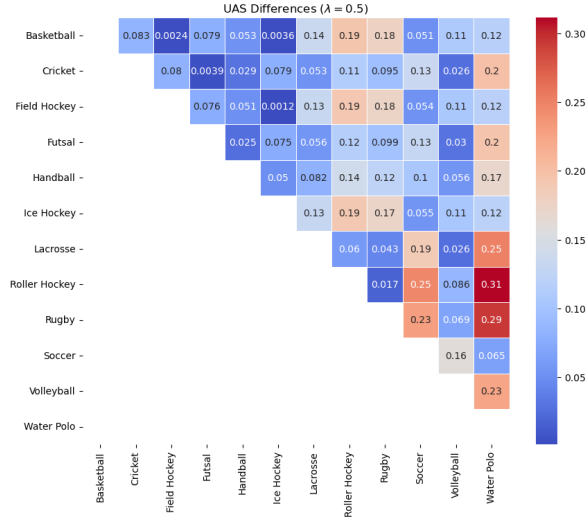
Figure 3: Upper triangular matrices representing the differences in UAS values for all pairs of sports, with the weight λ in (2.4) set to 0.5.

tends to be greater diversity, which increases the likelihood of weaker teams defeating stronger ones. A similar explanation applies to lacrosse, where UAS_{e_h} values remained at 0 prior to 1995, when fewer than seven teams participated in the World Lacrosse Men's World Cup, as indicated in [1, Table 7]. Aside from these observations, no other particular trends are evident in the graphs.

4 Randomness model

In this section, we develop a model consisting of randomness factors that can affect match outcomes in team ball sports. In Section 5, such a model will be used to gain insights into the relationship between the randomness factors and underdog achievement. Unlike [16] and [11], which propose variables of randomness affecting goal scoring in soccer as the match progresses, our model focuses on static factors, assuming scores as given. Therefore, we exclude factors that may influence player motivation, such as match location and current score, which are known to impact all sports and are not of interest to our analysis.

The factors that contribute to the inherent randomness observed in match outcomes are listed in Table 4 and categorized into three main groups: physical environment, player, and team. We believe that each of the factors in such a table should have a positive impact on randomness, meaning that a larger factor value corresponds to increased randomness. To provide further clarity into the relationship between such factors and randomness, we will include explanations when describing the three groups of factors in more detail. Although we believe the factors we have chosen comprehensively account for the observed variability in match outcomes across different sports, we acknowledge that the role of the factors in increasing or decreasing such variability reflects our own perspective. Table 5 includes companion factors used in the definition of some of the randomness factors in Table 4.

For each sport, we quantified the average values of the randomness factors, resulting in a factors dataset containing 12 rows (one for each sport) and 13 columns (one for each factor).
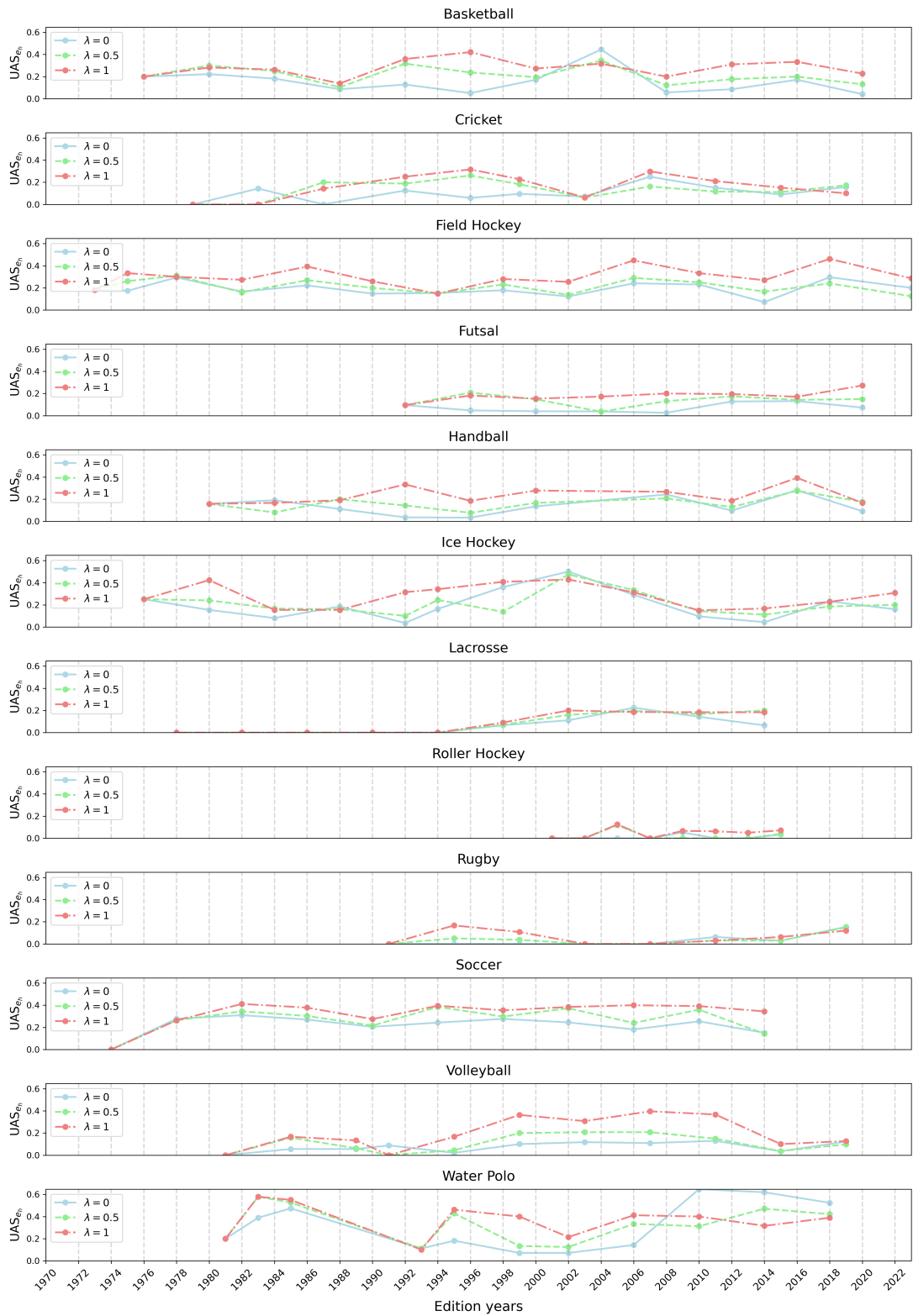
Figure 4: Graphs depicting the evolution of UAS_{e_h} over time for each sport for three different values of the weight λ in (2.4).

Such a factors dataset, presented in Table 7 of Appendix C, is provided in its normalized version. Table 8 serves as an auxiliary dataset used in the computation of values for Table 7. The sources for the values of the factors come from various websites and are available upon request. We will use the term "scoring targets" to collectively refer to the designated areas that allow teams to score, such as goals for soccer, baskets for basketball, and similar terms for the other team ball sports considered in our paper. The values in the factors dataset in Table 7 were derived by first using the formulas defined in the next subsections of this section and then applying normalization to rescale the range of each column in $[0, 1]$. When applying normalization to each column, we used the formula

$$a' = \frac{a - \min(a)}{\max(a) - \min(a)},$$

where a' denotes the normalized value, a denotes the original value, and $\min(a)$ and $\max(a)$ represent the minimum and maximum values that a takes on, respectively.

Physical Environment	
BL:	Ball lightness
BV:	Ball velocity
FS/BS:	Field size/Ball size
SS/BS:	Scoring target size/Ball size
BB:	Ball bounciness
Player	
PP:	Player powerfulness
PBH:	Player ball handling
PBD:	Player ball dispossession
PI:	Player inexperience
Team	
NP/FS:	Number of players/Field size
SS/NPS:	Scoring target size/Number of players who can effectively defend the scoring target
SI:	Scoring infrequency
NRAM/NRPM:	Number of rules about movement/Number of rules that prevent movement

Table 4: Factors that contribute to randomness in the match outcomes of the team ball sports included in our paper.

BW:	Ball weight
PBP:	Player ball possession
PE:	Player experience
SF:	Scoring frequency

Table 5: Companion factors used in the definition of some of the randomness factors in Table 4.

Physical environment factors. The physical environment category includes randomness factors related to properties of the sporting equipment and playing field. The formulas used to define such factors (including the units of measurement) are as follows:

BL $\max(BW) - BW$, where BW is the ball weight (gr)

BV   Average speed at which a player shoots the ball (km/h)

FS/BS   Surface of the field/Surface of the ball (m^2/m^2)

SS/BS   Surface of the scoring target/Surface of the ball (m^2/m^2)

BB   Categorical variable with eleven classes

            (from 0 for ice hockey to 11 for basketball).

Ball lightness (BL), which is inversely related to ball weight (BW), influences the force required for players to control a ball. We expect that lighter balls contribute more to randomness because they are generally more difficult to control due to their reduced mass, responding differently to player actions. Ball velocity (BV) affects the timing of the gameplay, with faster balls expected to increase randomness. The ratio between field size and ball size (FS/BS) is related to spatial dynamics. Higher values for such a ratio are associated with less ball control and more player movement, thus increasing randomness. The ratio between the scoring target size and ball size (SS/BS) influences the dynamics of a match in a similar way, as higher values for such a ratio imply a higher likelihood of achieving a scoring target and changing the match outcomes. Since cricket lacks a scoring target, we estimate its ratio between the scoring target size and ball size by averaging values obtained for other sports. Ball bounciness (BB), which determines the extent of rebound upon impact, was treated as a categorical variable by assigning each sport to one of eleven categories, from 0 for ice hockey (no bounciness) to 11 for basketball (maximum bounciness).

Player factors. The player category focuses on player attributes and skills that contribute to randomness. The formulas used to define such factors (including the units of measurement) are as follows:

PP   Body mass index = $\text{Weight}/\text{Height}^2$ (kg/m^2)

PBH   Proportion of body interacting with the ball

PBD $\max(PBP) - PBP$, where PBP is the player ball possession, defined as follows:

            PBP = Actual play time/Total number of players on the field (minutes/players)

PI $\max(PE) - PE$, where PE is the player experience, defined as follows:

            PE = Average retirement age (years).

Player powerfulness (PP) is related to the strength with which a player strikes a ball, influencing its trajectory and speed. Such a factor is measured in terms of the body mass index of a player, which is defined as the body mass divided by the square of the body height. Higher powerfulness is expected to increase randomness. Player ball handling (PBH) refers to the percentage of the body used to control a ball. In the case of cricket, lacrosse, field hockey, ice hockey, and roller hockey, such a percentage takes into account the sticks. Player ball dispossession (PBD) refers to a player's inability to maintain possession of a ball and, therefore, is inversely related to player ball possession (PBP), which we measure by the actual play time (i.e., match time, without including interruptions) divided by the total number of players on the field. Player ball

dispossession influences scoring opportunities because the lower the possession, the lower the control on a ball, and the higher the contribution to randomness. Player inexperience (PI) is inversely related to player experience (PE), which we measure in terms of average retirement age. The average retirement age reflects the accumulation of skills and decision-making abilities over time, and thus results in performance consistency. Therefore, as the level of inexperience increases, so does the contribution to randomness.

Team factors.   The team category includes randomness factors related to collective dynamics and match rules. The formulas used to define such factors (including the units of measurement) are as follows:

NP/FS	Total number of players on the field/Surface of the field (players/m^2)
SS/NPS	Surface of the scoring target divided by the number of players who can effectively defend the scoring target (m^2/players)
SI	max(SF) $-$ SF, where SF is the scoring frequency, defined as follows: SF = Number of scoring targets achieved or points being scored per team divided by the actual play time (scoring targets/min)
NRAM/NRPM	Ratio between the number of rules about movement and the number of rules that prevent movement, as defined in Table 9 of Appendix C.

The ratio between the number of players and the field size (NP/FS) is a measure of the coverage of the field by players. A higher value of such a ratio is associated with a wider range of offensive and defensive strategies and, therefore, is expected to have a positive impact on randomness. The ratio between the scoring target size and the number of players who can effectively defend the scoring target (SS/NPS) is a measure of the defensive weakness of a team. The fewer players defend the scoring target, the higher the variability in match outcomes. We estimate the value of such a ratio for cricket, which lacks a scoring target, by averaging the values obtained for the other sports. Scoring infrequency (SI) refers to how rarely scoring targets are achieved during a match. Such a factor is inversely related to the scoring frequency (SF), which we measure by the number of scoring targets achieved or points being scored per team divided by the actual play time. Sports with a low number of scoring targets achieved per match are more sensitive to randomness (in the sense of the final outcome of a match) and, therefore, the scoring infrequency can significantly impact the overall match outcome. The presence of team rules that restrict movement imposes tactical constraints, limiting team dynamics and playstyle. Fewer movement constraints can result in greater unpredictability for match outcomes. Table 9 of Appendix C includes all the rules considered for the computation of the ratio between the number of rules about movement and the number of rules that prevent movement (NRAM/NRPM).

5   Explaining the underdog achievement through our randomness model

In this section, we perform a principal component analysis (PCA) and a correlation analysis to gain insights into the relationship between the UAS computed for each sport in Section 3 (see Table 3) and the randomness factors introduced in Table 4 of Section 4 and quantified in the factors dataset in Table 7 of Appendix C. PCA is a statistical technique used to reduce

the dimensionality of a dataset by computing a linear combination of its column values (one can think of a column as a vector) while preserving the maximum proportion of variability from the original dataset [10]. PCA plots can be used to visually represent the results of the analysis through scatterplots, and these are the types of plots that we use to determine the relative importance of the randomness factors for each sport. A correlation analysis will then be performed to study the linear correlation between each pair of factors, including the underdog achievement.

Principal component analysis. By applying PCA, one can transform the factors dataset in Table 7, consisting of 13 columns, into a reduced dataset with as many columns as the number of principal components selected, where each principal component is obtained as a linear combination of the columns in the original factors dataset. Figure 5 represents the PCA plots for the first two principal components resulting from the application of PCA to the original factors dataset (upper plot) and the factors dataset with an additional column consisting of the UAS values obtained from Table 3 in Section 3 when $\lambda = 1$ (lower plot). The PCA plots in Figure 5 provide a graphical representation of the relationships between sports and factors captured by the first two principal components. In both PCA plots, the data points, represented as blue dots, are associated with the sports. We use different shades of blue to denote whether their UAS is high (dark blue), medium (medium blue), or low (light blue), based on the results described in Section 3. The coordinates of each data point are determined by the so-called scores, which are new variables associated with the columns in the reduced dataset. In general, the first two principal components are the ones that capture the maximum variance in a dataset. In our case, they are able to explain 57.13% of the variability in the original factors dataset, as shown in the scree plot in Figure 6 (specifically, the first principal component explains 31.29% and the second one explains 25.84%). In addition to the data points associated with the sports, the PCA plots include vectors, referred to as loadings and depicted as line segments, which represent the contribution of each original factor to the variability in the factors dataset explained by the principal components. Factors associated with loading vectors with similar directions and magnitudes have similar importance in explaining the variability. In the PCA plots in Figure 5, the magnitude of each loading vector has been doubled for improved clarity.

By examining the positions of the data points associated with the sports and the directions and magnitudes of the loading vectors in each of the PCA plots in Figure 5, one can observe the relative importance of the randomness factors for each sport.

Observation 1. In soccer, Player ball handling, Ball bounciness, and SS/NPS exhibit high values because soccer players are allowed to use various body parts to interact with the ball, the ball tends to be highly bouncy, and there is only one player who defends the scoring target. We note that although Number of players/Field size is close to soccer (in the sense of the plot), soccer actually has one of the smallest values for such a ratio. The close proximity of Number of players/Field size to soccer is perhaps due to the close proximity of soccer to volleyball in the plot, which has a high value for such a ratio.

Observation 2. For the hockey sports (i.e., field hockey, ice hockey, and roller hockey), one can observe that the predominant randomness factors are Player inexperience, Scoring infrequency, Field size/Ball size, Ball lightness, and Ball velocity. Indeed, in these sports, players retire at a relatively young age, scoring frequency is lower compared to other sports like basketball, ball sizes are small (resulting in a high Field size/Ball size value), ball weight is light, and ball velocity is high.
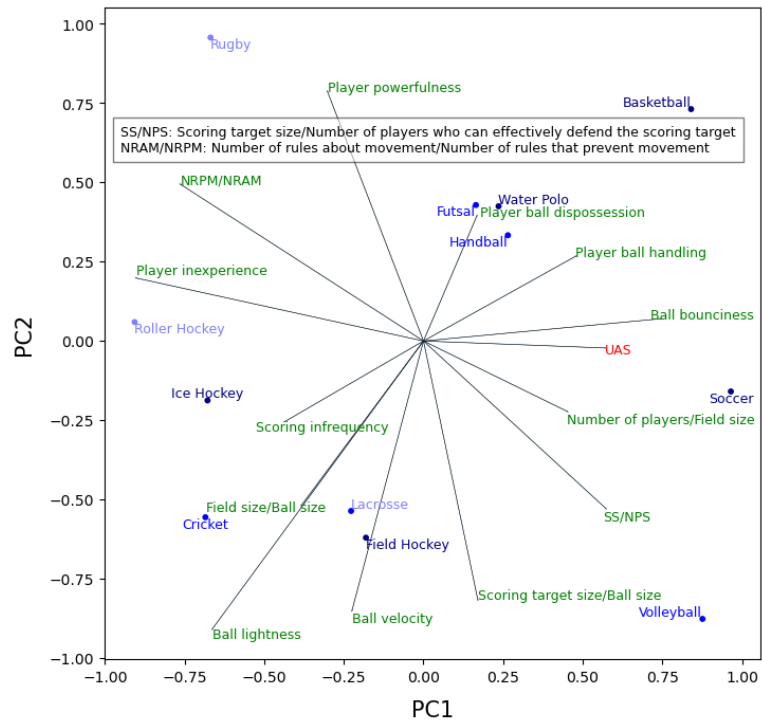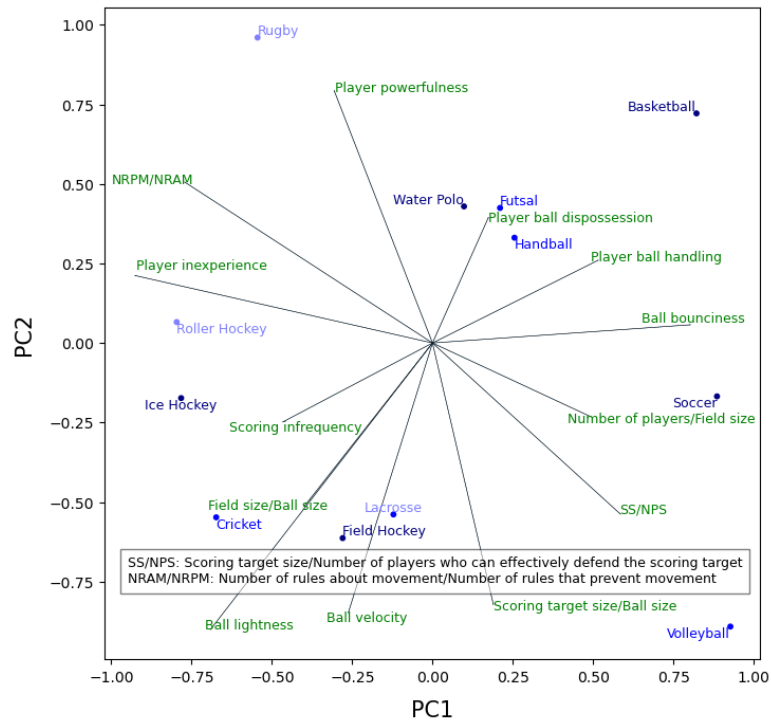
15

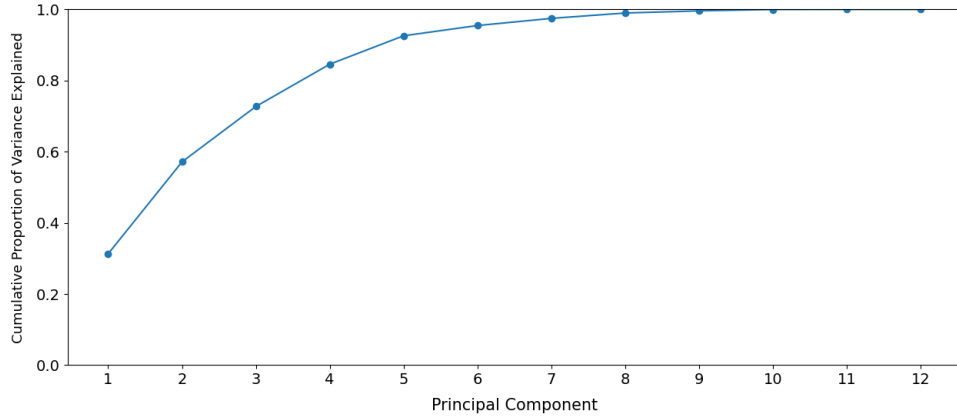Figure 5: PCA plots, without UAS (upper plot) and with UAS (lower plot).

Figure 6: Scree plot showing the cumulative proportion of variance explained by each additional principal component when UAS is not included among the factors. The scree plot when UAS is included is almost identical and is omitted.

Observation 3. For water polo, the main randomness factors are Player ball dispossession and Player powerfulness. The high Player ball dispossession value is due to the low actual play time compared to other sports. Similar conclusions can be extended to handball, futsal, and basketball, which are all close to water polo (in the sense of the plot). For basketball, the main randomness factors are Player ball dispossession, Player ball handling, and Ball bounciness.

Observation 4. For rugby, the most significant randomness factors are Player powerfulness, Player inexperience, and Player ball dispossession, each attaining their maximum values. Player inexperience has a high value because rugby players typically retire at a young age. The high Player ball dispossession value is due to a combination of very low actual play time and a very high number of players. NRAM/NRPM also takes on a high value due to the fewer movement restrictions in rugby compared to other sports.

Correlation analysis. To gain insights into the relationship between underdog achievement and randomness factors, we report Figure 7, which shows a heatmap illustrating the Pearson correlation coefficient between each pair of factors, including UAS. The Pearson correlation coefficient is a measure of the linear correlation between two variables, ranging from -1 (negative correlation) to 1 (positive correlation). The heatmap in Figure 7 indicates that UAS exhibits the strongest positive correlation with SS/NPS and the strongest negative correlation with NRAM/NRPM and Player inexperience. Additionally, UAS exhibits a relatively weaker negative correlation with Scoring infrequency, Player powerfulness, Field size/Ball size, and Ball lightness. The positive correlation between UAS and SS/NPS is expected, as SS/NPS takes on a high value in soccer, which is among the sports with the highest UAS. The negative correlations observed with UAS may appear surprising because they suggest that higher values of the factors decrease the underdog achievement, while we would expect that higher values of the factors increase randomness in match outcomes. However, this behavior is expected, as not all randomness factors have the same influence on underdog achievement. One can interpret the factors exhibiting a negative correlation with UAS as having a weaker effect on randomness than the other factors. For example, rugby has a high value for Player inexperience, which is negatively correlated with UAS. One can interpret this as the fact that the contribution of Player
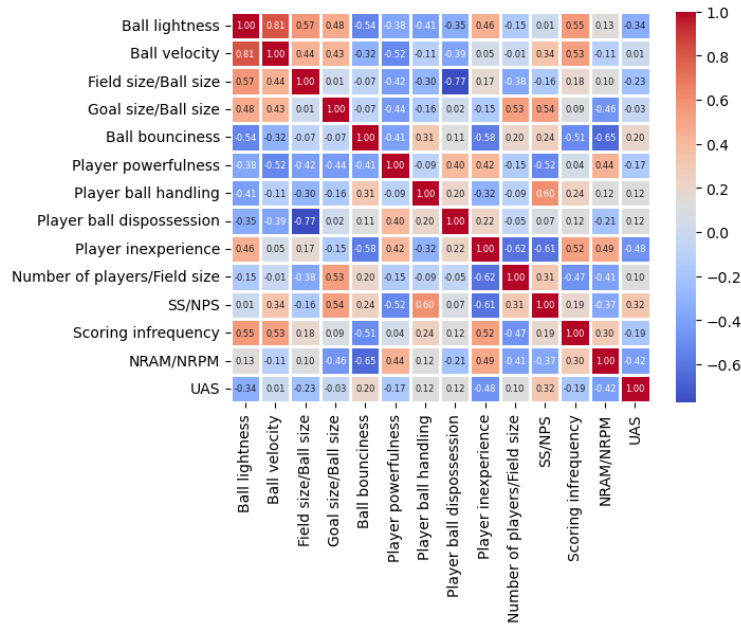
17

Figure 7: Heatmap illustrating the correlation between each pair of factors, including UAS.

inexperience to randomness in match outcomes is lower compared to other sports, which results in a low underdog achievement for rugby. By adopting such an interpretation, we conclude that the factors with the highest impact on randomness are those that exhibit a positive correlation with underdog achievement, i.e., SS/NPS, Number of players/Field size, Player ball dispossession, Player ball handling, Ball bounciness and, to a lesser extent, Scoring target size/Ball size, and Ball velocity.

6   Concluding remarks and future work

In this paper, we studied the relationship between underdog achievement and randomness factors that affect match outcomes in team ball sports. To achieve our goal, we collected match score data from major international competitions (World cups or Olympic games) held between 1970 and 2023, and we computed corresponding team rankings for each edition of the competitions selected for each sport. Then, we developed an underdog achievement score to determine the sports with the highest and lowest occurrences of weaker teams defeating stronger ones. Our findings indicate that water polo, soccer, field hockey, ice hockey, and basketball are among the sports with the highest underdog achievement, while lacrosse, roller hockey, and rugby are the ones with the lowest underdog achievement. Subsequently, we designed a randomness model consisting of 13 factors that contribute to unexpected match outcomes within each sport, providing quantitative values for each of the factors. Finally, we performed PCA and correlation analysis demonstrating that our randomness model can explain the underdog achievement. The randomness factors with the highest impact on underdog achievement are the ratio between the scoring target size and the number of players who can effectively defend the scoring target, the ratio between the number of players and the field size, player ball dispossession, player ball handling, and ball bounciness.

Although water polo appears to have a higher underdog achievement than soccer, our analysis confirms findings partially noted in the literature: soccer is among the sports with the highest underdog achievement, and we believe its inherent randomness is one of the reasons why soccer is universally recognized as the world's most popular sport [3]. Of course, other aspects contribute, such as the fact that playing soccer requires only a regular ball and no complex equipment. However, exploring such additional reasons is beyond the scope of our paper.

Our analysis has some limitations. As observed in Section 2, the match score dataset is a convenience sample, meaning it might not be fully representative of the entire population of sports matches. For futsal, roller hockey, and rugby, the data is limited either because their competitions are relatively new (futsal and rugby) or due to the unavailability of data from early editions (roller hockey). For water polo, data for some editions is missing, and for lacrosse, the early editions of the World Lacrosse Men's World Cup had fewer than seven teams (see [1] for more details). Due to the lack of access to official team rankings across different sports, we developed a team ranking logic that applies uniformly to all sports but does not account for the importance of specific matches within a competition, which can be relevant when assessing the strength of a team (e.g., strong teams may underperform or lose in less critical matches to conserve energy for more important matches, which does not necessarily indicate weakness). Finally, although we believe that the randomness factors we have chosen provide a reasonable explanation for the observed variability in match outcomes across different sports, we acknowledge that the role of the factors in increasing or decreasing such variability reflects our own perspective.

For future research, we plan to replicate the analysis by including women's sports competitions, competitions organized within professional sports leagues, and collegiate sports competitions. Additionally, we aim to investigate the applicability of the methodology to team non-ball sports.

Acknowledgments

References

[1] T. N. Alleck, T. Giovannelli, L. N. Vicente, R. Mitchell, and O. Remen. Match score dataset for team ball sports. *Data in Brief*, 55:110625, 2024.

[2] E. Ben-Naim, F. Vazquez, and S. Redner. Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2:1–1, 01 2006.

[3] J. Dvorak, A. Junge, T. Graf-Baumann, and L. Peterson. Football is the most popular sport worldwide. *The American journal of sports medicine*, 32:3S–4S, 01 2004.

[4] T. W. Edgar and D. O. Manz. Chapter 4 - exploratory study. In Thomas W. Edgar and David O. Manz, editors, *Research Methods for Cyber Security*, pages 95–130. Syngress, 2017.

[5] M. Elfil and A. Negida. Sampling methods in clinical research; an educational review. *Emerg (Tehran)*, 5:e52, 2017.

[6] B. Evangelos, A. Gioldasis, G. Ioannis, and A. Georgia. Relationship between time and goal scoring of european soccer teams with different league ranking. *Journal of Human Sport and Exercise*, 13:518–529, Sep. 2018.

[7] A. Galloway. Non-probability sampling. In Kimberly Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 859–864. Elsevier, New York, 2005.

[8] A. Heuer and O. Rubner. Fitness, chance, and myths: An objective view on soccer results. *The European Physical Journal B*, 67:445–458, 2008.

[9] L. M. Hvattum and H. Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26:460–470, 2010.

[10] I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 2016.

[11] M. Lames. Chance involvement in goal scoring in football – an empirical approach. *German Journal of Exercise and Sport Research*, 48:278–286, 2018.

[12] M. J. Lopez, G. J. Matthews, and B. S. Baumer. How often does the best team win? A unified approach to understanding randomness in north american sport. *The Annals of Applied Statistics*, 2017.

[13] M. Mauboussin. *The Success Equation: Untangling Skill and Luck in Business, Sports, and Investing.* Harvard Business Review Press, 2012.

[14] D. Sally and C. Anderson. *The Numbers Game: Why Everything You Know About Soccer Is Wrong.* Penguin Books Limited, 06 2013.

[15] H. Sarmento, F.M. Clemente, Afonso J., Araújo D., Fachada M., Nobre P., and Davids K. Match analysis in team ball sports: An umbrella review of systematic reviews and meta-analyses. *Sports Med Open*, 13, May 2022.

[16] F. Wunderlich, A. Seck, and D. Memmert. The influence of randomness on goals in football decreases over time. An empirical analysis of randomness involved in goal scoring in the english premier league. *Journal of Sports Sciences*, 39:2322–2337, 2021.

A   International competitions for the considered team ball sports

Table 6 contains the major international competitions and the corresponding edition years selected for the team ball sports included in our paper.
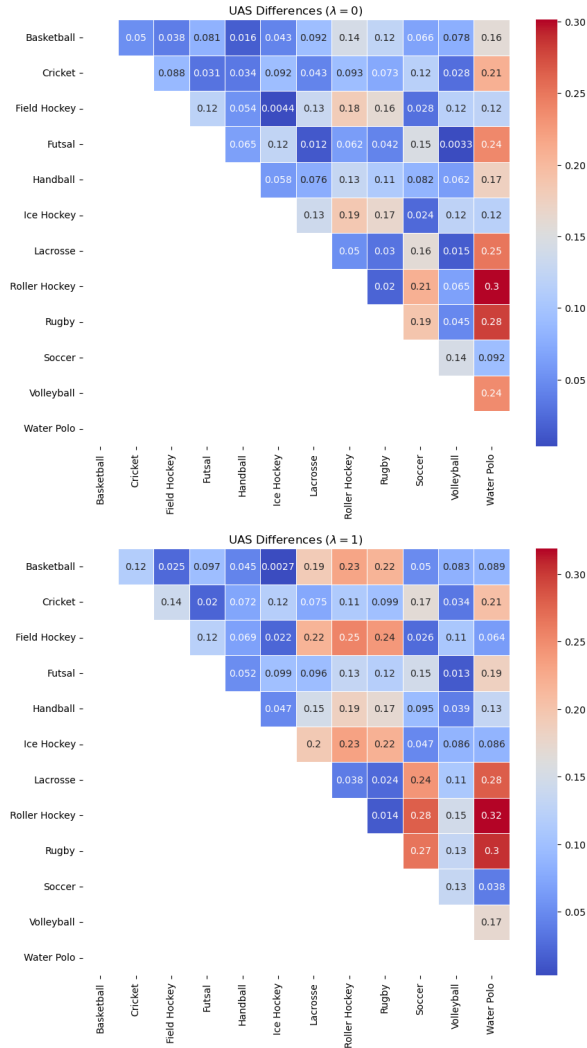
Figure 8: Upper triangular matrices representing the differences in UAS values for all pairs of sports when the weight λ in (2.4) is set to 0 and 1.

B   Upper triangular pairwise difference matrices ($\lambda \in \{0, 1\}$)

Figure 8 complements Figure 3 (which corresponds to $\lambda = 0.5$) by showing upper triangular matrices representing the differences in UAS values for all pairs of sports when the weight λ in (2.4) is set to 0 and 1.

C   Factors dataset

Table 7 is the factors dataset described in Section 4, provided in its normalized version. Table 8 serves as an auxiliary dataset used in the computation of values for Table 7. Table 9 summarizes the rules about movement and those preventing movement used in the computation of NRAM/NRPM in Table 8.

Sport	Major International Competition	Edition Years
Basketball	Summer Olympic Games	1972, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020
Cricket	ICC Men's Cricket World Cup	1975, 1979, 1983, 1987, 1992, 1996, 1999, 2003, 2007, 2011, 2015, 2019
Field Hockey	Men's FIH Hockey World Cup	1971, 1973, 1975, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014, 2018, 2023
Futsal	FIFA Futsal World Cup	1989, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020
Handball	Summer Olympic Games	1976, 1980, 1984, 1988, 1992, 1996, 2000, 2008, 2012, 2016, 2020
Ice Hockey	Winter Olympic Games	1972, 1976, 1980, 1984, 1988, 1992, 1994, 1998, 2002, 2006, 2010, 2014, 2018, 2022
Lacrosse	World Lacrosse Men's World Cup	1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
Roller Hockey	World Skate Roller Hockey World Cup	1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015
Rugby	Rugby World Cup	1987, 1991, 1995, 1999, 2003, 2007, 2011, 2015, 2019
Soccer	FIFA World Cup	1970, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
Volleyball	FIVB Volleyball Men's World Cup	1977, 1981, 1985, 1989, 1991, 1995, 1999, 2003, 2007, 2011, 2015, 2019
Water Polo	FINA Men's Water Polo World Cup	1979, 1981, 1983, 1985, 1993, 1995, 1999, 2002, 2006, 2010, 2014, 2018

Table 6: Major international competitions and corresponding edition years selected for the team ball sports included in our paper.

Sports	BL	BV	FS/BS	SS/BS	BG	BB	PP	PBH	PBD	PI	NP/FS	SS/NPS	SI	NRAM/NRPM
Basketball	0	0	0.0013	0	0.5	1	0.270	0.057	0.802	0.25	0.30	0	0	0.24
Cricket	0.97	0.80	1	0.066	0.5	0.4	0	0	0.52	0.55	0	0.073	0.83	0.75
Field Hockey	0.97	0.61	0.29	0.79	0.5	0.3	0.012	0	0.92	0.89	0.041	0.42	0.98	0.24
Futsal	0.40	0.41	0.0052	0.062	0.5	0.7	0.193	1	0.85	0.69	0.15	0.32	0.95	0.75
Handball	0.33	0.41	0.0061	0.074	0.5	0.7	0.431	0.057	0.94	0.60	0.22	0.024	0.82	0
Ice Hockey	0.96	1	0.11	0.23	0	0	0.486	0	0.79	0.91	0.079	0.10	0.99	tcb0.59
Lacrosse	1	0.74	0.46	0.43	0.5	0.9	0.012	0	0.94	0.92	0.031	0.17	0.92	0
Roller Hockey	0.98	0.59	0.056	0.164	0.5	0.1	0.259	0	0.82	0.92	0.12	0.077	0.93	1
Rugby	0.37	0.089	0.042	0.126	1	0.2	1	0.32	0	1	0.029	0.039	0.93	0.94
Soccer	0.38	0.67	0.070	0.188	0.5	0.8	0.009	1	0.91	0.083	0.0096	1	1	0.24
Volleyball	0.73	0.74	0	1	0.5	0.6	0.020	0.19	0.65	0	1	0.75	0.59	0
Water Polo	0.39	0.35	0.0013	0.0095	0.5	0.5	0.463	0.057	0.94	0.5	0.30	0.13	0.77	0.42

Table 7: Normalized factors dataset that quantifies the values of the randomness factors described in Section 4 for each of the team ball sports included in our paper.

Sports	BW (g)	BV (km/h)	FS/BS (m²/m²)	SS/BS (m²/m²)	BG	BB	PP (kg/m²)	PBH	PBP (min/pl.)	PE (years)	NP/FS (pl./m²)	SS/NPS (m²/pl.)	SF (targets/min)	NRAM/NRPM
Basketball	602	29	$2.4 \cdot 10^3$	119.1	1	10	24.78	0.092	4.8	33	0.023	0.45	2.4	1.29
Cricket	159.5	128	$1.0 \cdot 10^6$	NA	1	4	23.15	0.054	19.09	29.39	0.0014	NA	NA	1.6
Field Hockey	160	104	$3. \cdot 10^5$	460.6	1	3	23.22	0.054	2.73	25.37	0.0044	7.83	0.073	1.29
Futsal	420	80	$6.3 \cdot 10^3$	47.6	1	7	24.31	0.72	4	27.76	0.012	6	0.15	1.6
Handball	450	79.2	$7.2 \cdot 10^3$	54.1	1	7	25.74	0.092	2.29	28.8	0.018	0.86	0.46	1.14
Ice Hockey	163	152.5	$1.1 \cdot 10^5$	144	0	0	26.08	0.054	5	25.1	0.0071	2.16	0.12	1.5
Lacrosse	145	121	$4.6 \cdot 10^5$	257.6	1	9	23.22	0.054	2.4	25	0.0036	3.35	0.21	1.14
Roller Hockey	155	102	$5.7 \cdot 10^4$	105	1	1	24.71	0.054	4.5	25	0.01	1.79	0.2	1.75
Rugby	435	40	$4.3 \cdot 10^4$	84	2	2	29.17	0.27	1.27	24	0.0035	1.12	0.2	1.71
Soccer	430	112	$7.1 \cdot 10^4$	119.1	1	8	23.20	0.72	2.91	35	0.0021	17.86	0.03	1.29
Volleyball	270	121	$1.2 \cdot 10^3$	578.6	1	6	23.27	0.18	7.5	36	0.074	13.5	1	1.14
Water Polo	425	72	$2.5 \cdot 10^3$	17.8	1	5	25.93	0.092	2.29	30	0.023	2.7	0.56	1.4

Table 8: Auxiliary dataset used in the computation of values for Table 7. For quantities that are not adimensional, we include the corresponding unit of measurement below the name of the column.

Sport	Rules about movement (RAM) and rules preventing movement (RPM)
Basketball (RAM: 1–9) (RPM: 1–7)	1. Traveling: Restricts players from taking steps without dribbling the ball. 2. Dribbling: Governs the legal handling of the ball while moving. 3. Charging and blocking: Regulates contact between offensive and defensive players. 4. Impeding: Restricts obstructing an opponent's movement without playing the ball. 5. 3-Second violation: Limits offensive players' time spent in the key area. 6. 5-Second violation: Restricts the time allowed to inbound or shoot free throws. 7. Out-of-bounds: Determines possession and restarts play when the ball goes out of bounds. 8. Illegal contact: Penalizes players for illegal physical contact with opponents. 9. Offensive foul for pushing off: Prohibits offensive players from pushing off.
Cricket (RAM: 1–8) (RPM: 1–5)	1. Running between the wickets: Regulates the movement of batsmen between wickets. 2. Crease and stump movement: Regulates player positioning near the crease and stumps. 3. Fielding position restrictions: Specifies fielding positions and limitations during play. 4. Fielding the ball: Regulates the legal method of fielding and returning the ball. 5. Running in the protected area: Prohibits running in areas designated for protection. 6. Backfoot no-ball rule: Penalizes bowlers for overstepping the crease during delivery. 7. Fair and unfair play: Governs fair play conduct and penalizes unfair actions on the field. 8. Pitch etiquette: Specifies behavior and conduct on the pitch during play.
Field Hockey (RAM: 1–9) (RPM: 1–7)	1. Offside: Regulates player positioning relative to the opponents during play. 2. Advantage: Allows play to continue for minor infractions. 3. Obstruction: Prohibits players from blocking opponents' access to the ball. 4. Dangerous play: Penalizes players for actions that may endanger themselves or others. 5. Pushing: Regulates the legal method of using the stick to push the ball. 6. Impeding: Prohibits players from obstructing opponents' movement without the ball. 7. Illegal contact: Penalizes players for illegal physical contact with opponents. 8. Back stick: Penalizes players for using the rounded backside of the stick to play the ball. 9. High-sticking: Prohibits players from using sticks above shoulder height.
Futsal (RAM: 1–8) (RPM: 1–5)	1. Running with the ball: Regulates dribbling and movement with the ball. 2. 3-Second rule: Limits the time a player can hold the ball without dribbling or passing. 3. Goalkeeper restrictions: Specifies rules and limitations unique to the goalkeeper position. 4. Encroachment on free kicks: Prohibits players from encroaching during free kicks. 5. Kicking in: Governs the method of restarting play from the touchline. 6. Intentional time wasting: Penalizes teams for deliberately wasting time during play. 7. Five-foul limit: Penalizes teams for committing a certain number of fouls in a half. 8. No slide tackling: Prohibits slide tackling to ensure player safety.
Handball (RAM: 1–8) (RPM: 1–7)	1. Dribbling: Regulates the movement of the ball while in hand possession. 2. 3-Second violation: Limits the time a player can hold the ball without passing or shooting. 3. Stepping inside the goal area: Limits players from entering the goal area during play. 4. Jumping: Regulates jumping actions, particularly during shooting or passing. 5. Encroachment on free throws: Restricts opponents from entering the free-throw area. 6. Goalkeeper: Specifies actions and limitations unique to the goalkeeper position. 7. Holding and pushing: Penalizes players for holding or pushing opponents illegally. 8. Impeding: Restricts obstructing an opponent's movement without playing the ball.
Ice Hockey (RAM: 1–9) (RPM: 1–6)	1. Offside: Prevents attacking players from entering the offensive zone before the puck. 2. Impeding: Prohibits obstructing or interfering with an opponent without the puck. 3. Hooking: Restricts players from using their stick to hook an opponent. 4. Slashing: Penalizes players for swinging their stick at an opponent. 5. Boarding: Penalizes players for checking an opponent into the boards violently. 6. Charging: Penalizes players for charging into an opponent violently. 7. Icing: Regulates the clearing of the puck from one end of the rink to the other. 8. Tripping: Penalizes players for causing opponents to fall by tripping them. 9. High-sticking: Prohibits players from using sticks above shoulder height.
Lacrosse (RAM: 1–8)	1. Offside: Regulates player positioning on the field during play. 2. Moving picks or screens: Prohibits illegal screens to obstruct defenders.

(RPM: 1–7)	3. Crease violation: Penalizes players for entering the crease area during play.
	4. Stick checks: Regulates legal stick checking actions during play.
	5. Offensive fouls: Penalizes offensive players for illegal actions during play.
	6. Illegal picks: Prohibits illegal screens set by offensive players.
	7. Impeding: Prohibits obstructing or interfering with an opponent without the ball.
	8. Illegal contact: Penalizes players for illegal physical contact with opponents.
Roller Hockey (RAM: 1–7) (RPM: 1–4)	1. Offside: Regulates player positioning relative to the puck during play.
	2. Obstruction: Prohibits obstructing opponents without playing the puck.
	3. Crease violation: Penalizes players for entering the crease area during play.
	4. Tripping: Penalizes players for causing opponents to fall by tripping them.
	5. Illegal contact: Penalizes players for illegal physical contact with opponents.
	6. Illegal screen: Prohibits illegal screens to obstruct opponents.
	7. High-sticking: Prohibits contacting the puck with the stick above shoulder height.
Rugby (RAM: 1–12) (RPM: 1–7)	1. Offside: Regulates player positioning during set plays and general play.
	2. Advantage: Allows play to continue for minor infractions.
	3. Maul: Governs the formation and legality of mauls during play.
	4. Scrum: Regulates the engagement and conduct of scrums to restart play.
	5. Lineout: Specifies rules and procedures for lineouts to restart play from touch.
	6. Kicking: Regulates kicking actions during open play and set pieces.
	7. Not retreating 10 meters: Penalizes teams for insufficient retreat from penalties.
	8. Tackling: Governs legal tackling techniques and player safety during tackles.
	9. Ruck: Regulates player actions and roles in rucks formed during play.
	10. Illegal blocking or holding: Prohibits illegal blocking or holding actions during play.
	11. Dangerous play: Penalizes players for actions that endanger opponents' safety.
	12. Not binding correctly: Regulates the correct binding of players in scrums and mauls.
Soccer (RAM: 1–9) (RPM: 1–7)	1. Offside: Prevents forward players from receiving the ball behind the defense.
	2. Handling: Prohibits using hands or arms to control the ball, except for the goalkeeper.
	3. Goalkeeper: Specifies actions and limitations unique to the goalkeeper position.
	4. Impeding: Restricts obstructing an opponent's movement without playing the ball.
	5. Fouls against goalkeepers: Protects goalkeepers from physical contact.
	6. Charging opponents: Prohibits excessive or dangerous body contact with opponents.
	7. Impeding opponents: Prevents deliberately impeding an opponent's progress on the field.
	8. Fouls and misconduct: Governs various rule violations, including fouls and misconduct.
	9. Simulation and diving: Penalizes players for simulating fouls or exaggerating contact.
Volleyball (RAM: 1–8) (RPM: 1–7)	1. Rotational faults: Regulates player positions during serve and rotation.
	2. Foot faults during service: Prevents foot faults during the serving motion.
	3. Back-row attack violations: Limits back-row players from attacking past the 3-meter line.
	4. Blocking across the net: Governs legal blocking actions across the net.
	5. Center line violations: Prohibits players from crossing the center line during play.
	6. Libero restrictions: Specifies actions and limitations for the libero player.
	7. Illegal substitutions: Regulates player substitutions and entry onto the court.
	8. Net violations: Penalizes players for touching the net during play.
Water Polo (RAM: 1–7) (RPM: 1–5)	1. Holding or pushing: Prohibits holding or pushing opponents in the water.
	2. Impeding: Prohibits obstructing or interfering with an opponent without the ball.
	3. Sinking: Prohibits players from deliberately sinking or diving to gain advantage.
	4. Kick-off: Specifies rules and procedures for the kick-off to start the game.
	5. Corner throw: Governs the method of restarting play from the corner of the pool.
	6. Striking: Penalizes players for striking opponents with hands or arms.
	7. Two-handed pushoff: Regulates the legal use of hands to push off opponents.

Table 9: Table that summarizes the rules about movement (RAM) and those preventing movement (RPM) used in the computation of NRAM/NRPM in Table 8.