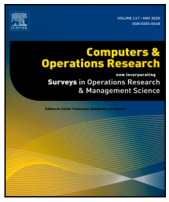
Computers & Operations Research 144 (2022) 105821

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor



The efficacy of tournament designs

Balázs R. Sziklai [a,c], Péter Biró [a,c], László Csató [b,c,*]

[a] Centre for Economic and Regional Studies (KRTK), Budapest, Hungary
[b] Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network (ELKH), Laboratory on Engineering and Management Intelligence, Research Group of Operations Research and Decision Systems, Budapest, Hungary
[c] Corvinus University of Budapest (BCE), Department of Operations Research and Actuarial Sciences, Budapest, Hungary

ARTICLE INFO

Keywords:
Competitive balance
OR in sports
Ranking
Simulation
Tournament design

ABSTRACT

Tournaments are a widely used mechanism to rank alternatives in a noisy environment. This paper investigates a fundamental issue of economics in tournament design: what is the best usage of limited resources, that is, how should the alternatives be compared pairwise to best approximate their true but latent ranking. We consider various formats including knockout tournaments, multi-stage championships consisting of round-robin groups followed by single elimination, and the Swiss-system. They are evaluated via Monte-Carlo simulations under six different assumptions on winning probabilities. Comparing the same pair of alternatives multiple times turns out to be an inefficacious policy. While seeding can increase the efficacy of the knockout and group-based designs, its influence remains marginal unless one has an unrealistically good estimation on the true ranking of the players. The Swiss-system is found to be the most accurate among all these tournament formats, especially in its ability to rank all participants. A possible explanation is that it does not eliminate a player after a single loss, while it takes the history of the comparisons into account. The results can be especially interesting for emerging esports, where the tournament designs are not yet solidified.

1. Introduction

We study the following *ranking problem*. There is a set of alternatives ("players"), characterised by a single attribute ("strength"). The decision-maker does not know the true strengths but can observe the outcome of paired comparisons ("matches") between any two players. The results of these clashes are noisy in the sense that a stronger player does not always defeat a weaker one, however, the winning probability monotonically increases as a function of the difference in ability. The aim is to rank *all* players according to their strengths as well as possible. However, we do not consider ties and home advantage.

The schedule of paired comparisons is called *tournament format*. The economics, management science, and sports literature mostly discuss the situation when the players choose the intensity of their effort, and the principal's objective is to provide incentives for achieving its goal(s) (Lazear and Rosen, 1981; Rosen, 1986; Taylor, 1995; Prendergast, 1999; Szymanski, 2003; Orrison et al., 2004; Brown and Minor, 2014; Bimpikis et al., 2019). On the contrary, here we consider the level of effort to be fixed: it is assumed that all players always perform at their best. This can be a realistic hypothesis in many high-stake environments like crowdsourcing contests (Hou and Zhang, 2021), elections (Klumpp and Polborn, 2006), innovation races (Harris and Vickers, 1987; Yücesan, 2013; Ales et al., 2017), musical competitions (Ginsburgh and Van Ours, 2003), or sports tournaments (Palacios-Huerta and Volij, 2009).

There are two basic tournament formats. In the binary, knockout, or single-elimination (henceforth knockout) tournament, the loser of any match is immediately eliminated and cannot be the winner. The selection efficiency of this design has been extensively discussed in economics and statistics, especially concerning the effects of its seeding procedure (Hartigan, 1968; Israel, 1981; Hwang, 1982; Horen and Riezman, 1985; Knuth and Lossers, 1987; Chen and Hwang, 1988; Edwards, 1998; Schwenk, 2000; Glickman, 2008; Vu and Shoham, 2011; Groh et al., 2012; Prince et al., 2013; Kräkel, 2014; Hennessy and Glickman, 2016; Karpov, 2016; Adler et al., 2017; Dagaev and Suzdaltsev, 2018; Karpov, 2018; Arlegi and Dimitrov, 2020; Kulhanek and Ponomarenko, 2020; Arlegi, 2021). The second prominent format is the round-robin, in which all players face all the others and the players are ranked according to their results (Harary and Moser, 1966; Rubinstein, 1980).

Efficacy can be defined as the capability of a tournament to reproduce the ranking of the players according to their strength. In the real-world, this ranking is naturally hidden, although there are good

* Corresponding author at: Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network (ELKH), Laboratory on Engineering and Management Intelligence, Research Group of Operations Research and Decision Systems, Budapest, Hungary.
E-mail address: laszlo.csato@sztaki.hu (L. Csató).

https://doi.org/10.1016/j.cor.2022.105821
Received 7 July 2021; Received in revised form 17 February 2022; Accepted 29 March 2022
Available online 6 April 2022

proxy measures (e.g. Elo scores) based on past performance. Here we assume that the power ranking is known and the matches are decided accordingly.

The exact probability p_{ij} of player i finishing in place j can be derived for a small number of players (David, 1959; Glenn, 1960; Searls, 1963). However, this approach becomes impossible for tournaments with a large number of matches and a complicated branch structure. For instance, there are $2^{n(n-1)/2}$ possible outcomes in a round-robin tournament with n players.

Therefore, we estimate the probabilities by a Monte Carlo simulation, which is a standard methodology in the literature. Appleton (1995) aims to determine the chance that the best player wins a particular competition, including random and seeded knockout, round-robin, draw and process, and Swiss-system. McGarry and Schutz (1997) reveal the ranking efficacy of some traditional tournament structures for eight players under a variety of initial conditions. Mendonça and Raghavachari (2000) compare multiple round-robin tournament ranking methods with respect to their ability to replicate the true rank order of players' strengths. Marchand (2002) computes the chances of a top-seeded player winning a standard and a random knockout tournament, and gives evidence that the outcome of the two "antagonistic" versions may not vary as much as expected. According to Ryvkin and Ortmann (2008), the predictive power – the probability of selecting the best player as the winner – of knockout and round-robin tournaments exhibits non-monotonicity as a function of the number of players for fat-tailed distributions of abilities. Ryvkin (2010) explores two alternative measures of selection efficiency of these mechanisms, the expected ability of the winner and the expected rank of the winner.

Further studies approach the problem from the perspective of sports. Scarf et al. (2009) provide a comprehensive overview of tournament formats used in practice and present an extensive list of metrics, as well as a simulation framework. Scarf and Yusof (2011) continue this analysis by examining the effect of seeding policy on outcome uncertainty. Goossens et al. (2012) compare four league formats that have been considered by the Belgian Football Association. Annis and Wu (2006) assess the relative merits of playoff scenarios for NCAA I-A football, which differ in the number, selection, and seeding of playoff teams. Lasek and Gagolewski (2018) investigate the efficacy of league formats used in the majority of European top-tier association football competitions. Csató (2021b) analyses four hybrid designs, consisting of knockout and round-robin stages, used in the recent IHF World Men's Handball Championships.

Generally, the simulations reinforce the statistical principle that a larger sample (more matches played) leads to better estimates (Lasek and Gagolewski, 2018; Csató, 2021b). However, none of these works have addressed explicitly a fundamental issue of economics: what is the best usage of limited resources, i.e. which format should be followed if the aim is to approximate the true ranking with a given number of costly matches. Furthermore, except for Appleton (1995), previous works have not examined the Swiss-system.

It is important to note here that the choice of tournament format is driven by a variety of factors such as fairness (Cea et al., 2020; Chater et al., 2021; Durán et al., 2017; Goossens and Spieksma, 2012; Guyon, 2015, 2018, 2020; Kendall et al., 2010; Kendall and Lenten, 2017; Laliena and López, 2019; Van Bulck and Goossens, 2020; Wright, 2014), incentive compatibility (Csató, 2020, 2021a, 2022; Dagaev and Sonin, 2018; Pauly, 2014; Preston and Szymanski, 2003; Vong, 2017), maximising attendance (Krumer, 2020), or minimising rest mismatches (Atan and Çavdaroğlu, 2018). Nonetheless, accuracy in the ranking of the competitors is clearly among the most important aspects of tournament design since in several sports, broadcasting revenues are distributed on the basis of the teams' final position in the ranking (Bergantiños and Moreno-Ternero, 2020; Petróczy and Csató, 2021).

Appleton (1995) and McGarry and Schutz (1997) have studied tournaments with eight or 16 players. This paper, similarly to Scarf et al. (2009), analyses a tournament with 32 competitors. The enlargement allows considering a broader range of structures, e.g. multi-stage tournament with eight groups. Since the number of competitors in several contests is larger than 16, this modification strengthens the applicability of our results, too.

Our major contribution resides in showing that the Swiss-system is basically more efficacious than any other tournament format containing the same number of matches, especially with respect to its ability to reproduce the *full* ranking of the players. On the other hand, increasing the number of matches between the same players seems to be wasteful: it is better to integrate two separate knockout contests in an ingenious way than to organise more matches according to the same schedule. Seeding can substantially improve the efficacy of a format, however, it requires an unrealistically good prediction about the true ranking of the players. Real data suggest that actual performance in a tournament can significantly differ from the past performance of the players, leaving seeding a significant, but ultimately marginal element in increasing the accuracy of a format. This is reinforced by a recent statistical study, which reveals that seeding itself does not contribute positively to the success of the teams in the UEFA Champions League and the UEFA Europa League (Engist et al., 2021).

2. Methodology

We test the efficacy of various tournament designs via Monte Carlo simulations.

2.1. Tournament formats

Deriving a full ranking of the players raises two difficulties. First, certain formats, such as the knockout tournament, do not give a complete order, at least in their traditional configuration. Second, all ties should be resolved. The first issue is handled by organising extra matches between the players who are already eliminated, thus each competitor plays the same number of matches in each tournament format. For tie-breaking purposes, the traditional Sonneborn–Berger and Buchholz rules are applied. The following designs are implemented.

Round-robin: Each player plays one match with all other players. Ties are resolved by the Sonneborn–Berger score.

Double round-robin: Players participate in two round-robin tournaments. Ranking is derived from the combined scores, ties are broken again by the Sonneborn–Berger rule.

Knockout: In each round, players are paired to play a match. The loser is eliminated, while the winner proceeds to the next round. The process is repeated until a sole winner remains, which requires n rounds for 2^n players. In our simulation, the eliminated players also enter into a knockout tournament. That is, all players eliminated in the first round continue the competition and will be ranked between the 17th and 32nd places. Analogously, players eliminated in the second round compete further for the places from the 9th to the 16th, and so on. Consequently, each player plays five matches.

Triple knockout: A knockout system where elimination/progression is decided on the basis of three matches instead of only one. That is, the players are paired in each round to play three matches and the player with two wins qualifies for the next round—the third game is played even if it is unnecessary when the same player wins both the first and the second games.

Draw and process: It is a double elimination tournament. The players play two parallel knockout championships, seeded such that any players who clash in the first or second rounds of the first tournament (draw) cannot meet in the second tournament (proceed) until the final or semi-final, respectively. The final ranking is obtained by comparing the outcome of the two parts as follows.

Suppose that the first k positions of the final ranking have already been determined by looking at the first m positions of the knockout results, thus the player(s) ranked at the $(m+1)$th place in the knockout

tournaments are considered. If the same player occupies these positions, it will be the $(k+1)$th in the final ranking. If two different players occupy these positions, there are three different cases. If both of them have already obtained a rank in the final ranking, the investigation continues with the player(s) ranked $(m+2)$th. If exactly one of them has already obtained a rank in the final ranking, the other one will be the $(k+1)$th. Otherwise, if none of them have already obtained a rank, they play a tiebreaker match. The winner will occupy the $(k+1)$th position and the loser will be the $(k+2)$th in the final ranking.

This format is applied e.g. in croquet (Croquet Association, 2021, Chapter F1d: Two-Life Events).

Multi-stage tournament with 8 groups: The players are divided into eight groups of four players each, where they play a round-robin tournament. The top two players with the highest scores from each group advance to a knockout tournament to allocate the first 16 places in the upper bracket, while the bottom two from each group play a knockout tournament in the lower bracket. Ties in the round-robin phase are broken by the Sonneborn–Berger rule.

Multi-stage tournament with 4 groups: Analogous to the previous format, but now there are four groups composed of eight players each. Again, the upper and the bottom half of the players form two knockout tournaments.

Double group: Similarly to the multi-stage tournament with four groups, the players are divided into four groups of eight players each, where they play a round-robin tournament. The top four players from each group proceed to the second round robin-phase. Here four four-player groups are formed by taking the best player from one group, the second best from another, the third and the fourth best from the third and fourth group, respectively. That process is repeated three more times. In the end, the four newly formed groups each contain a winner, a runner-up, a third- and a fourth-placed player from the first round-robin phase. In parallel, the losers of the first round-robin phase form another four four-player groups (losing branch). In both round-robin phases, ties are resolved by the Sonneborn–Berger rule. The final ranking is derived by a knockout tournament that follows the round-robin phases. Winners and second place runner-ups of the second round-robin compete for the first 8th position. Third- and fourth-place players compete for the 9–16th positions. The 17–24th and 25–32th positions are determined by the matches of the losing branch.

Swiss-system: It is a non-eliminating tournament format with a fixed number of rounds. The matching algorithm pairs players with (approximately) the same score in every round but two players cannot meet more than once (Biró et al., 2017; Führlich et al., 2021). In our simulation, the pairing is based on an integer program that aims to match the players with the highest score first. In particular, the program finds a maximum weight matching on a graph where the nodes represent the players and two nodes are connected if the corresponding players have not played against each other. The weights of the edges are the product of the players' incremented score (current score +1). The +1 increment is needed to avoid zero weights for players who have not yet won any match. The program yields the same result as the blossom algorithm developed by Jack Edmond.

The winner is the player having the most wins at the end. Ties are resolved after the final round by the Buchholz scores, which sums the scores of all opponents. In contrast to the previous designs, the Swiss-system depends on one parameter, the number of rounds.

This format is commonly used in bridge, chess (FIDE, 2020, Chapter C04: FIDE Swiss Rules), croquet (Croquet Association, 2021, Chapter F3: Swiss Events), Go, and esports (Blizzard Entertainment, 2020).

2.2. Match prediction and simulation

A symmetric probability matrix is used: for each pair of players A and B, the fixed winning probability $p_{AB} = 1 - p_{BA}$ determines the likelihood that player A wins against player B. The ranking is transitive, furthermore, if $p_{AB} > 0.5$ and $p_{BC} > 0.5$, then $p_{AC} \geq \max\{p_{AB}; p_{BC}\}$.

Appleton (1995) and Mendonça and Raghavachari (2000) use normally distributed ratings, whereas McGarry and Schutz (1997) consider fixed, linearly structured values. Since the number of players is smaller in these articles (eight or 16 in Appleton (1995), six in Mendonça and Raghavachari (2000), and eight in McGarry and Schutz (1997)), their assumptions on winning probabilities cannot be uniquely extended to our case. Scarf et al. (2009) use historical data from the UEFA Champions League to predict the outcome of matches, thus their simulation model depends on the sport considered to a large extent. Therefore, we have decided to examine three theoretical and three real-world scenarios, consequently, all simulations are carried out with six different sets of winning probabilities.

For each scenario, we conducted 1 million simulation runs. The metrics presented in the following are derived from these simulations. According to the Central Limit Theorem, the sample means converge. To show that the convergence is sufficiently fast, we cut the 1 million simulation runs into 10 equal sized runs and tested the obtained sample means with Student's t-test. At the significance level of 1%, the sample mean of each metric presented here does not differ from the true value by more than 0.5%.

In the three theoretical scenarios, the probability that player A wins against player B is given by

$$P_{AB} = 0.5 + \texttt{skill} \times (\texttt{rank(B)} - \texttt{rank(A)})/100,$$

where `skill` is a parameter and `rank(A)` and `rank(B)` denote the positions of the two players in the real power ranking, respectively. In the first theoretical scenario, even the weakest player has a reasonable chance to win against the strongest player (`skill` = 1). In the third, only top players can compete with the best (`skill` = 10). The middle models a transition between these two extremities (`skill` = 5).

For instance, if A is the strongest player and B is the second one under `skill` = 10, then the former wins with a probability of 60%. On the other hand, the winning probability of player A is reduced to 51% if `skill` = 1. If rank difference would indicate a negative probability, it is treated as an impossible event. Similarly, if rank difference leads to a winning probability greater than 1, it is considered to be a sure event.

We also consider real data from three competitions in different sports where the performance of the players can be measured by the Elo scores (note that this methodology is used "officially" only in the case of chess):

- Chess: the players of the World Chess Cup 2017;
- Soccer: the clubs participating in the 2017/18 UEFA Champions League;
- Tennis: the contestants of the 2017 Monte-Carlo Rolex Masters -- Singles tournament.

Elo data are obtained from https://en.wikipedia.org/wiki/Chess_World_Cup_2017 for chess, http://clubelo.com/ for soccer (retrieved on 12th September 2017), and http://tennisabstract.com/ for tennis (retrieved 9th April 2017), respectively. This is a convenient approach for our model because the Elo rating system is explicitly designed to reflect the winning probabilities of the players against each other. Since draws are not allowed, the winning probability of player A against player B is given by the following formula:

$$P_{AB} = \frac{1}{1 + 10^{(R_B - R_A)/400}},$$

where R_A and R_B are the Elo rating of players A and B, respectively.

Heatmaps of the winning probabilities in the six scenarios are provided in the Appendix, see Table A.1.

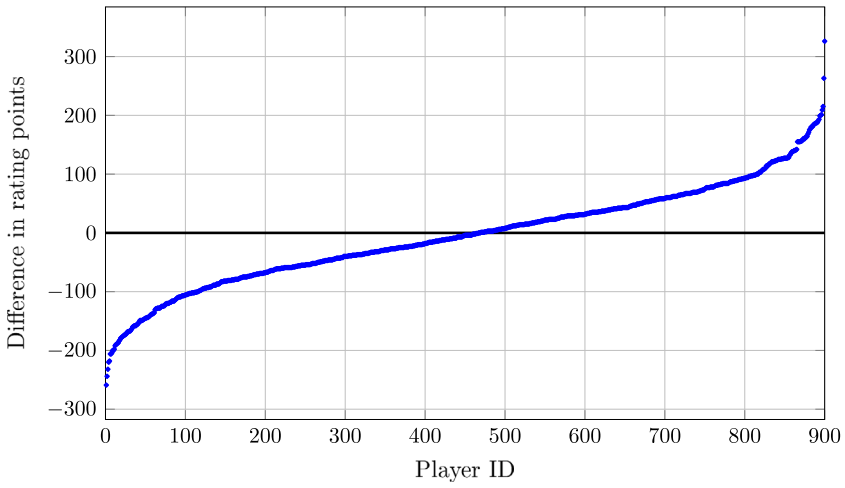


Fig. 1. Difference between the initial rating of the players and their performance.

Table 1
Examples for the calculation of the weighted inversion metric.

Reference ranking	Ranking A	Ranking B	Ranking C
1	2	1	3
2	1	2	5
3	3	3	1
4	4	5	4
5	5	4	2

Grey cells highlight players that precede their reference rank.

Table 2
The number of matches in the tournament formats with 32 players.

Tournament format	Number of matches
Round-robin	496
Double round-robin	992
Knockout	80
Triple knockout	240
Draw and process	160
Multi-stage with 8 groups	112
Multi-stage with 4 groups	176
Double group	208
Swiss-system	16 × number of rounds

2.3. Tournament metrics

Efficacy is quantified by looking at the differences between the real and observed rankings. For this purpose, two types of indicators are used, the average rank of the top players and the number of (weighted) inversions. The former has been applied by Scarf et al. (2009). It compares the top k players in the observed ranking to the k strongest players by dividing the sum of the ranks of the players finishing in the top k positions with the theoretical minimum of $1+2+\cdots+k = k(k+1)/2$.

However, if $k = 3$, this metric does not differentiate between the rankings $3 \succ 2 \succ 1$ (when the third best player wins the tournament) and $1 \succ 2 \succ 3$ (when the strongest player wins the tournament). Consequently, the number of inversions is also computed: the number of times when a weaker player is ranked above a stronger one. In the previous example, the number of inversions is equal to 3 for the ranking $3 \succ 2 \succ 1$, but it is 0 for the ranking $1 \succ 2 \succ 3$.

Furthermore, people tend to care more about the winners, and the top places of the rankings attract more attention, hence differences from the real power ranking are more noticeable here (Can, 2014). Therefore, we also propose a novel weighted inversion metric that sums the reciprocals of the logarithms of disconcordant positions.

Table 1 presents three examples. A player with reference rank j that is ranked at position $i < j$ adds $\sum_{k=i+1}^{j} 1/\ln(k)$ to the value of the metric. In ranking A, the only player who is ranked higher than its reference rank is player 2. Since the inversion takes place in the first position, the reciprocals of the logarithms are summed up starting from $i+1 = 2$ to $j = 2$, which is just a one-term sum. On the other hand, two players precede their reference rank in ranking C. Player 3 adds $1/\ln 2 + 1/\ln 3$ weight, while Player 5 adds $1/\ln 3 + 1/\ln 4 + 1/\ln 5$. Thus the weighted inversions between the three rankings are as follows:

$$w(A) = \frac{1}{\ln 2};$$
$$w(B) = \frac{1}{\ln 5};$$
$$w(C) = \frac{1}{\ln 2} + \frac{1}{\ln 3} + \frac{1}{\ln 3} + \frac{1}{\ln 4} + \frac{1}{\ln 5}.$$

Note that both rankings A and B contain only one inversion compared to the reference ranking, however, in ranking A this happens to be in the top position, hence it weighs more. This is manifested in $w(A) = 1/\ln 2 > 1/\ln 5 = w(B)$.

These measures characterise the efficacy of tournament formats sufficiently well. A lower value of them is always preferred to a higher one.

Table 3
The probability that the Swiss-system with five rounds leads to a ranking having a smaller number of inversions than the knockout.

Model	Probability
`skill = 1`	0.6350
`skill = 5`	0.9044
`skill = 10`	0.9379
Chess	0.5598
Soccer	0.6676
Tennis	0.7095

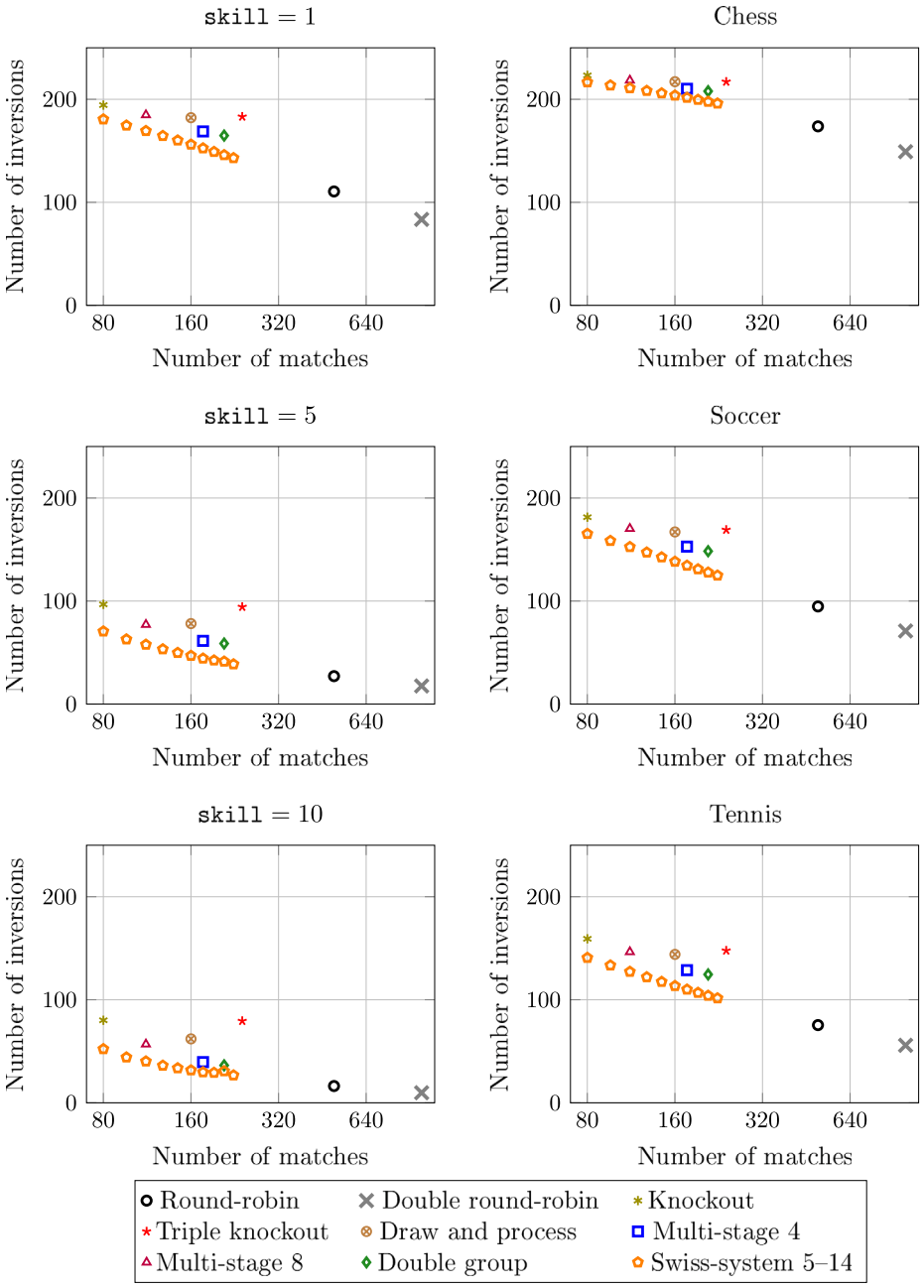


Fig. 2. The connection between the number of inversions and the number of matches (logarithmic scale on the horizontal axis).

2.4. Seeding

To disentangle the effect of seeding from tournament structure, random seeding is considered as our baseline. In each simulation run, a new random order of the players is generated.

The impact of seeding is investigated in two ways. To uncover the maximum possible effect of seeding, the true ranking of the players is assumed to be known and they are seeded according to this ranking. Therefore, the seeding is the same in each simulation run.

In particular, the standard seeding is used for the knockout, triple knockout, and draw and process formats. This seeding method is often used in practice and has been extensively studied in the literature (see, for example, Hwang (1982) or Schwenk (2000)). Analogously, the traditional method is applied for the group stage in any tournament design. If there are k groups and gk teams, the k highest-ranked players are placed in pot 1, the next k in pot 2, and so on, hence pot g contains the k lowest-ranked players. After that, each group gets one player from each pot randomly.

Finally, we estimate the expected impact of seeding. Before the players are seeded, their power ranking is perturbed because the organisers do not know the true ranking, they only have a guess based on the previous matches of the players.

In chess, tournament performances are routinely measured since rating performance can be used for tie-breaking. The results of six chess tournaments were taken into account:

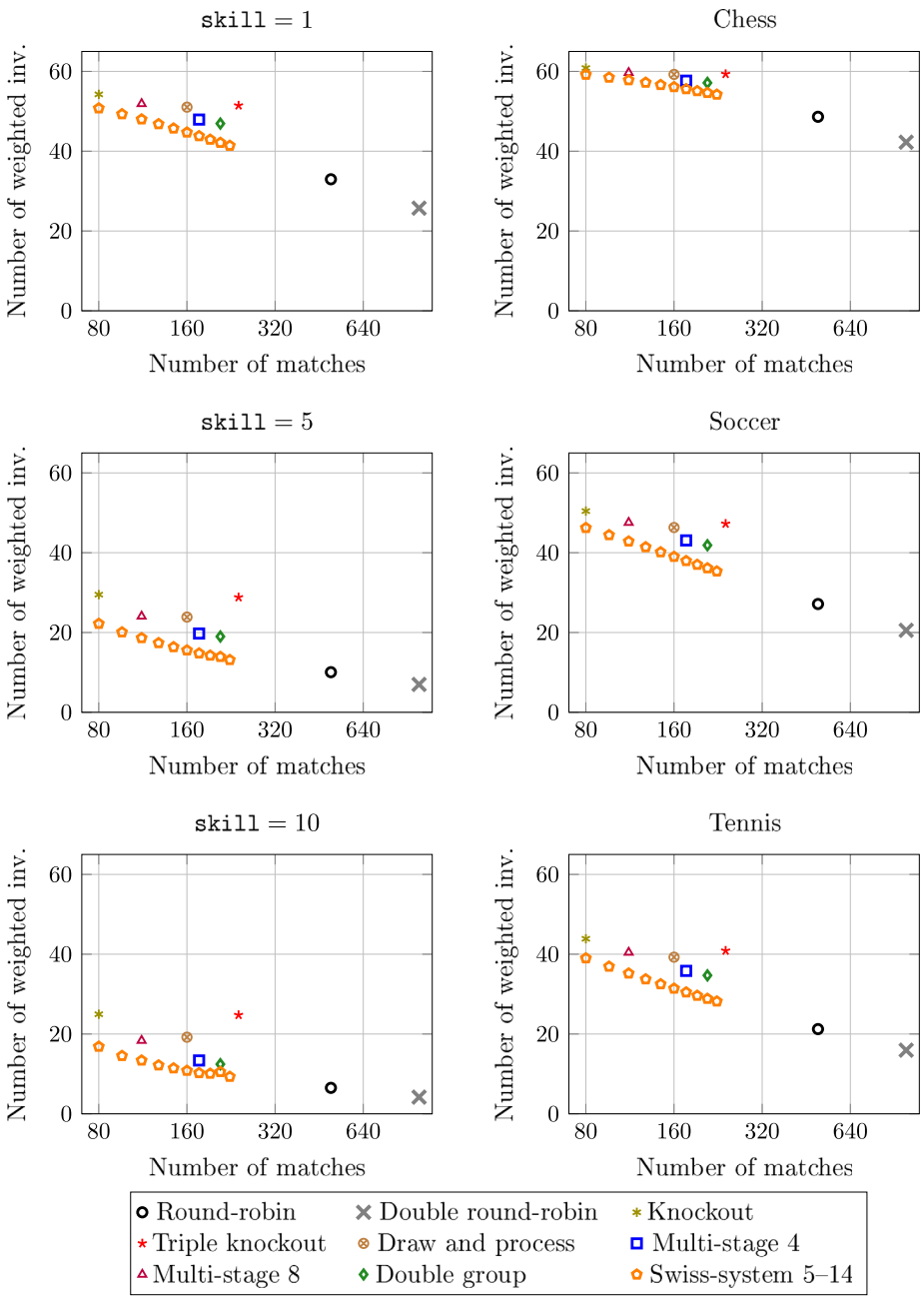


Fig. 3. The connection between the number of weighted inversions and the number of matches (logarithmic scale on the horizontal axis).

- European Individual Chess Championship 2017;
- European Individual Chess Championship 2018;
- European Individual Chess Championship 2019;
- Isle of Man 2017 Open – Masters;
- Gibraltar International Chess Festival 2019 – Masters;
- Grand Swiss 2019.

Data was gathered from the website . All six events were Swiss-system tournaments with 9, 10, or 11 rounds. For each tournament, the performances of the top 150 players are considered, leading to 900 data points. The majority of the players were grand masters (GM). For each tournament, the initial rating and the rating performances of the players were used: their difference indicates how the actual and past performances differ. Tournament performance was calculated according to the FIDE regulations.

Fig. 1 presents the result. As expected, approximately half of the players underperformed (overperformed) in the tournaments. Furthermore, the difference between the initial and the realised Elo rating is below (above) 100 points for about one-ninth of the players. To calculate the expected impact of seeding, we start from the real Elo data in the chess scenario. In each simulation run, performance differences are drawn randomly from the empirical distribution given in Fig. 1. The perturbed ratings represent the real power ranking, while the preliminary ranking is the initial rating used for seeding.

Seeding is not considered for the round-robin and double round-robin designs as each player plays the same number of matches against all the others. Analogously, seeding is not studied for the Swiss-system,

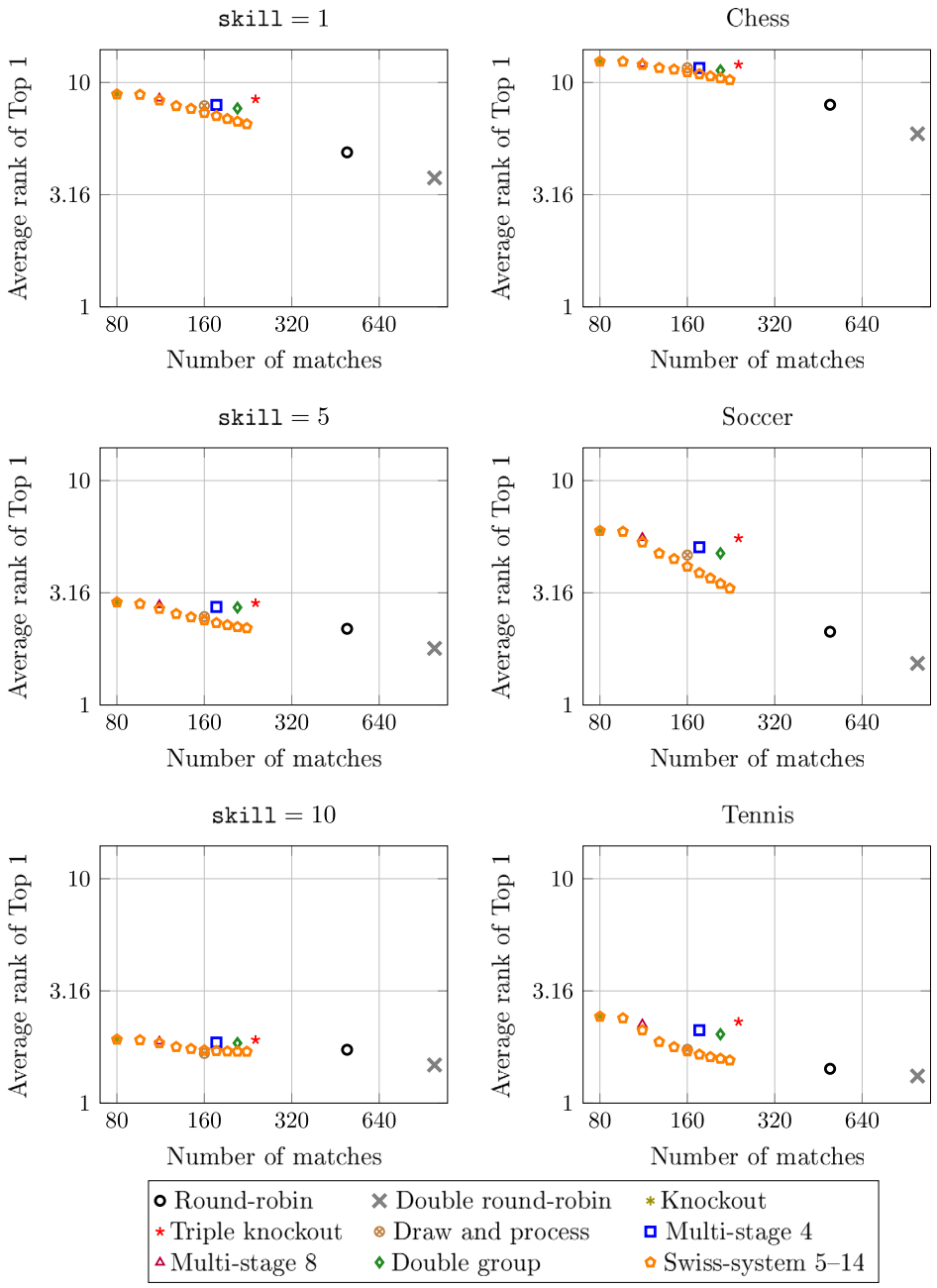


Fig. 4. The connection between the average rank of Top 1 and the number of matches (logarithmic scale on both axes).

where the pairing in each round is determined by the results of the previous rounds, except for the first.

3. Results

Our discussion begins with the analysis of unseeded tournament formats, followed by examining the effect of seeding and comparing the findings to previous results.

3.1. Random seeding

According to Table 2, the number of matches substantially differs for the tournament designs considered. However, the strength of the players can be estimated only by playing matches, thus the efficacy of any ranking mechanism highly depends on the number of matches played. This is shown in Figs. 2–5 by four measures, the number of inversions, the number of weighted inversions, as well as the average rank of the winner (Top 1) and the first eight players (Top 8). For the Swiss-system, the number of rounds is varied between 5 and 14 to contain at least as many matches as the knockout tournament.

The analysis is centred around two issues: (1) What is the best way to increase the number of matches if efficacy should be improved? (2) How should the competition be designed for a given number of matches?

Concerning the first question, we compare the knockout and round-robin structures, together with the draw and process that consists of

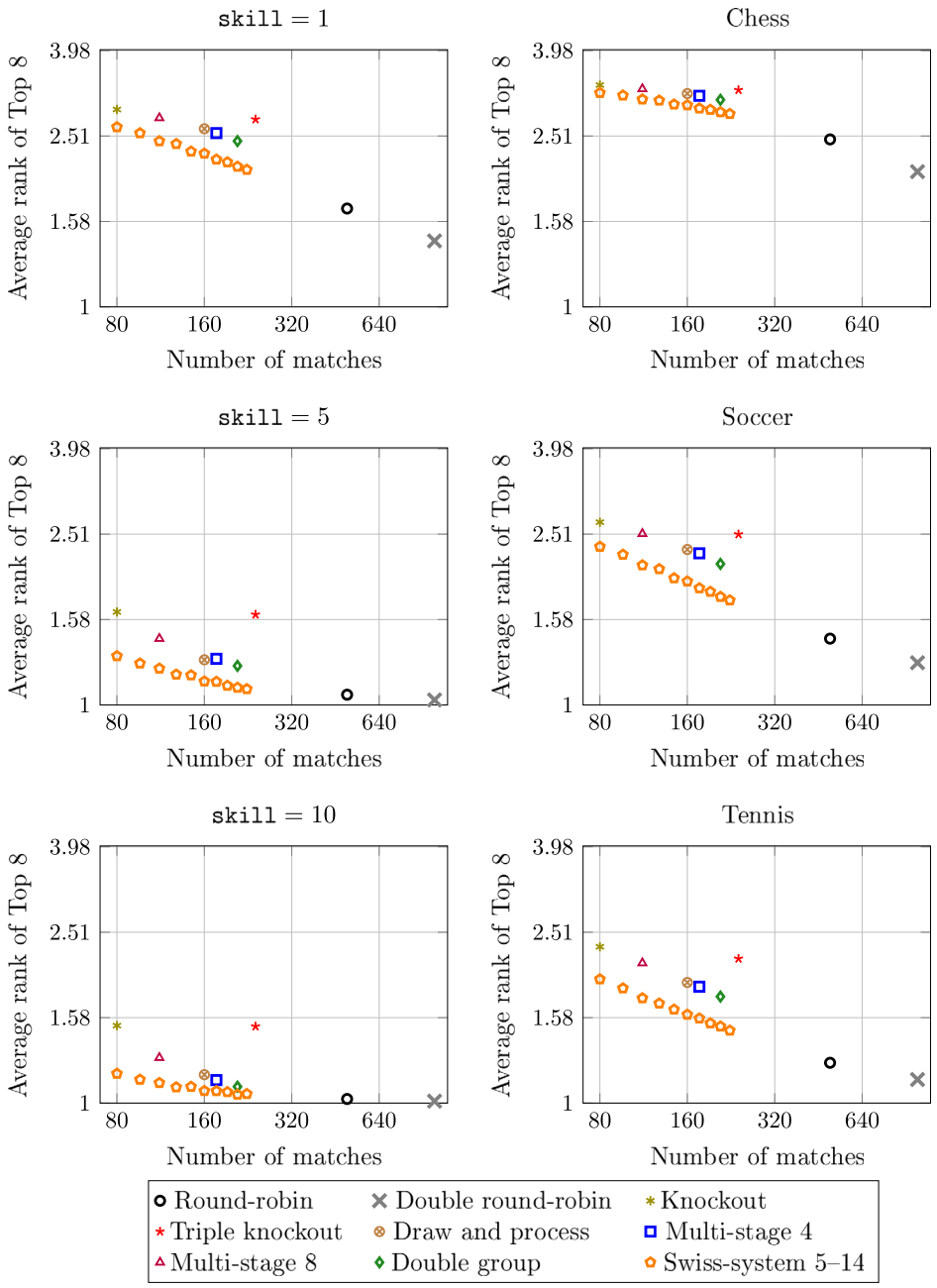


Fig. 5. The connection between the average rank of Top 8 and the number of matches (logarithmic scale on both axes).

essentially two knockout contests. As expected, the two round-robin formats outperform the others. The gain from playing two round-robin championships instead of one is higher when the competition is less balanced: the number of inversions is reduced by approximately 15% for our chess data, 25% for soccer and tennis, as well as for `skill` = 1, but 40% if `skill` = 10.

Using a triple knockout system rather than a simple one is clearly uneconomical, the maximal gain in the number of inversions is less than 8% under all probability models. Integrating two knockout competitions according to the draw and process system is robustly more efficacious than playing three matches for elimination. Interestingly, the relative advantage of draw and process over knockout does not change for its ability to select the winner, which is somewhat counter-intuitive since the latter mechanism is explicitly designed for this purpose.

To conclude, increasing the number of matches in the same design is not a parsimonious way to improve efficacy. Nonetheless, the possibility of ties and home advantage are disregarded in the analysis. Since these factors might be significant (for example, in professional tennis, see Koning (2011)), their incorporation offers an interesting line of future research.

In order to answer the second question, all formats except for the round-robins can be compared to a Swiss-system tournament with the

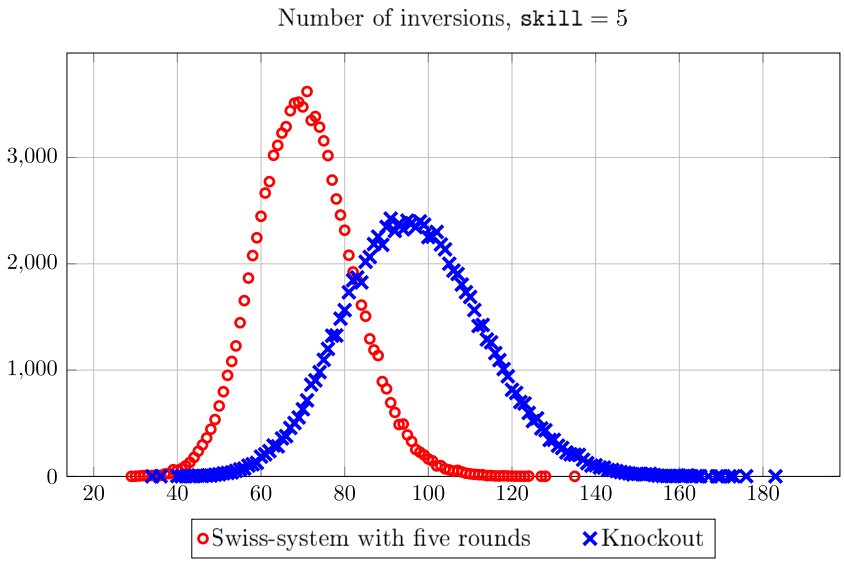


Fig. 6. The distribution of a tournament metric for two formats.

same number of matches, see Table 2. In each round of a Swiss-type tournament with 32 players, there are 16 matches. Hence, a 5-round Swiss-tournament has 80 matches just like the simple knockout, whereas a 7-round Swiss competition contains 112 matches, similar to the multi-stage tournament with eight groups.

The corresponding Swiss-system is robustly preferred to any group-based tournament (multi-stage with 4 or 8 groups, double group). In the case of real data, the advantage of the Swiss-system is found to be higher if the competition is less balanced, that is, in tennis. The gain from the Swiss-system compared to a more simple design is the lowest in chess, which can be surprising because the Swiss-system is applied in this particular sport, showing the strength of traditions. The tournament measures of the Swiss-system converge to the corresponding measures of round-robin since the latter is equivalent to a Swiss-system where the number of rounds is the number of players minus one.

Note that the metrics of the Swiss-system are non-monotonic as a function of all matches played if $\texttt{skill} = 10$. The ability of the round-robin tournament to select the winner is unexpectedly weak in the view of how the Swiss-system performs under $\texttt{skill} = 5$ and $\texttt{skill} = 10$. Finally, the draw and process design is competitive against the Swiss-system if the aim is to determine a sole winner, especially for $\texttt{skill} = 10$ when the winning probability of the stronger player often equals one (under this assumption, the best player always wins against the players ranked 6–32). The same observation holds for the knockout format. A possible explanation is that the best player can suffer an unexpected loss, thus playing more matches is favourable for reproducing the full ranking but not necessarily for selecting the winner.

These results are based only on averages. Fig. 6 presents the discrete distribution of the number of inversions for the knockout format and the Swiss-system with five rounds in the model $\texttt{skill} = 5$ (based on 100 thousand simulation runs). Since the distribution can be well approximated by the normal distribution, the averages reliably describe the efficacy of the tournaments.

Table 3 reports the derived estimations that the Swiss-system with five rounds outperforms the knockout design with respect to number of inversions. The advantage of the Swiss-system increases as the skill differences become more pronounced, which reinforces the message of Figs. 2–5.

The variance of the metrics can uncover how consistent the formats are in finding the true ranking. Fig. 7 shows the relative standard deviation of the number of inversions. The tournament designs do not differ much in this aspect but there are a couple of interesting observations. Firstly, the variance is quite high, especially when the skill differences are significant: in the $\texttt{skill} = 10$ scenario, the relative standard deviation of the double round-robin format reaches almost 30%. Secondly, increasing the number of matches slightly increases the variance. In other words, one gets more accurate but somewhat less precise results. Nevertheless, Fig. 7 suggests that the averages in Figs. 2–5 reliably reflect the efficacy of the tournament formats.

3.2. Seeded tournaments

Fig. 8 shows that seeding can improve the efficacy of a tournament design up to 10% (see the formats named "Perfect" in Fig. 8), but only when there is reliable information on the true ranking of the players. Under realistic circumstances, when the true ranking has to be approximated, the gain is reduced to at most 3%, thus the Swiss-system remains more efficacious than any other design (see the formats named "Seeded" in Fig. 8). In both cases, seeding has the highest effect on the multi-stage tournament with 8 groups. On the other hand, draw and process becomes less efficacious with seeding. The likely explanation is that the second, process branch is designed according to seeding-like information, thus seeding in the first, draw branch leads to an inferior pairing in the process branch. Finding a reasonable seeding mechanism for this tournament format remains an interesting open question.

3.3. Comparison with previous results

Finally, it is worth comparing our results to earlier works in the literature. Appleton (1995) focuses on the percentage of times when the best player wins and finds that the Swiss-system does not perform well in this respect, for instance, draw and process seems to be better. According to Fig. 4, this does not hold in general, the Swiss-system is not dominated by the draw and process even in the ability to determine the best player. Analogously, the triple knockout format strongly improves the chances of the best player compared to the knockout if there are only eight players (Appleton, 1995) but the gain is moderated in our case (Fig. 4).

Similar to us, McGarry and Schutz (1997) notice that the knockout is a weak tournament in its ability to rank all players but its efficacy can

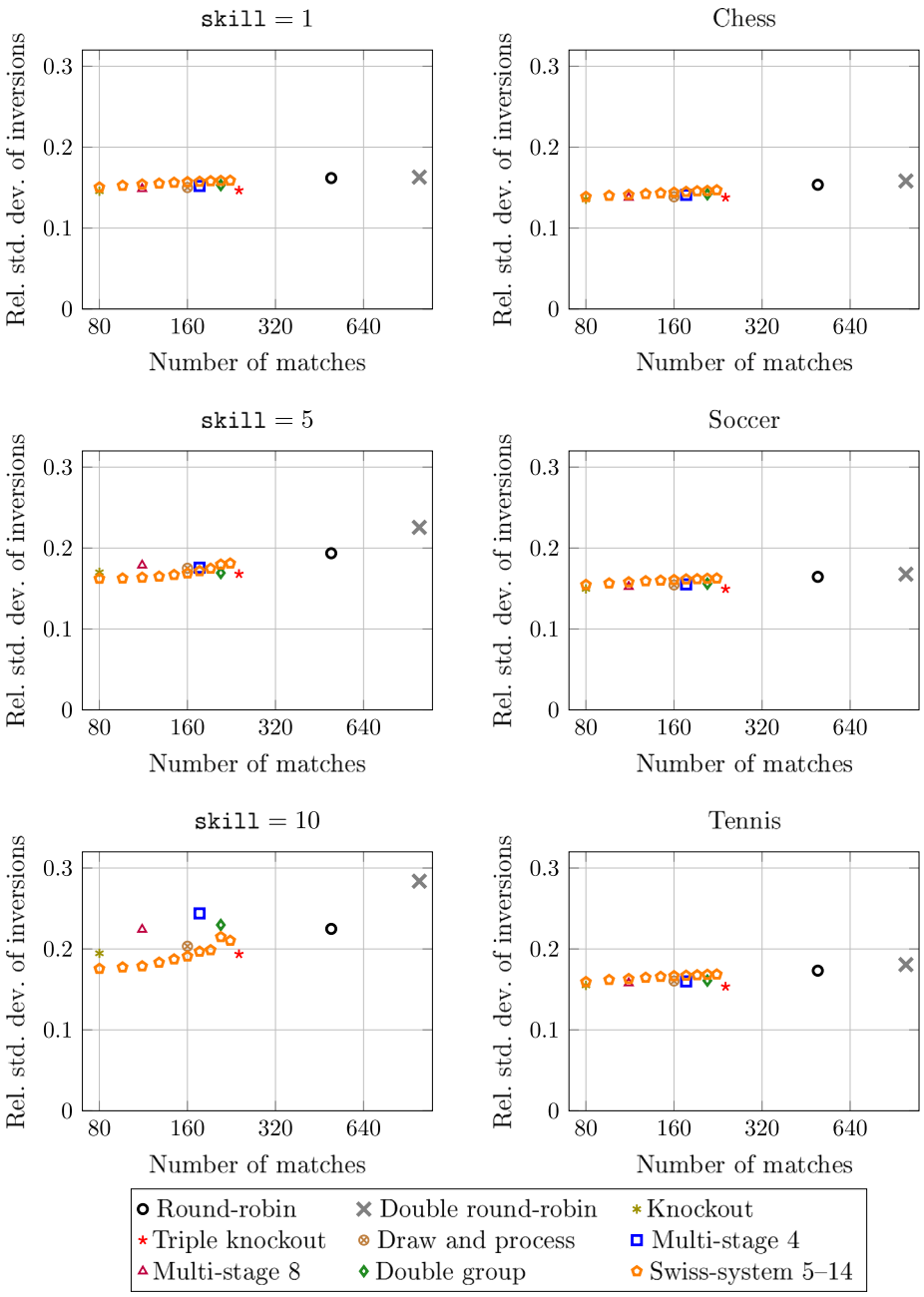


Fig. 7. The connection between the relative standard deviation of the number of inversions and the number of matches (logarithmic scale on the horizontal axis).

be enhanced by double elimination (draw and process), and especially, with an *accurate* seeding. However, seeding does not help much if the initial ranking differs from the true ranking as Fig. 8 reveals.

We also reinforce the results of Scarf et al. (2009): tournaments with group rounds imply a higher correlation of the pre-tournament to exit ranks than knockout tournaments but this has to be traded off against the number of matches. Therefore, the conclusions of Scarf et al. (2009) probably hold for a much larger set of winning probabilities. To conclude, while our results do not contradict the main findings of previous studies, we refine important details, especially with the incorporation of the Swiss-system into the analysis.

The lessons of our study can be summarised as follows: (1) the Swiss-system is more efficacious than any other tournament formats containing the same number of matches, with the possible exception of the average rank of the winner (Top 1) when its performance is similar to the knockout and draw and process designs, especially if one has an accurate seeding; (2) in the ability to reproduce the true ranking, the superiority of the Swiss-system increases with the number of matches taken into consideration; (3) draw and process, composed from two knockout modules, outperforms multi-stage tournaments for selecting the best player but not in the number of (weighted) inversions; (4) triple knockout shows a poor performance compared to the high number of matches played. All of the above findings are robust with respect to the derivation of the winning probabilities and the tournament metric chosen.

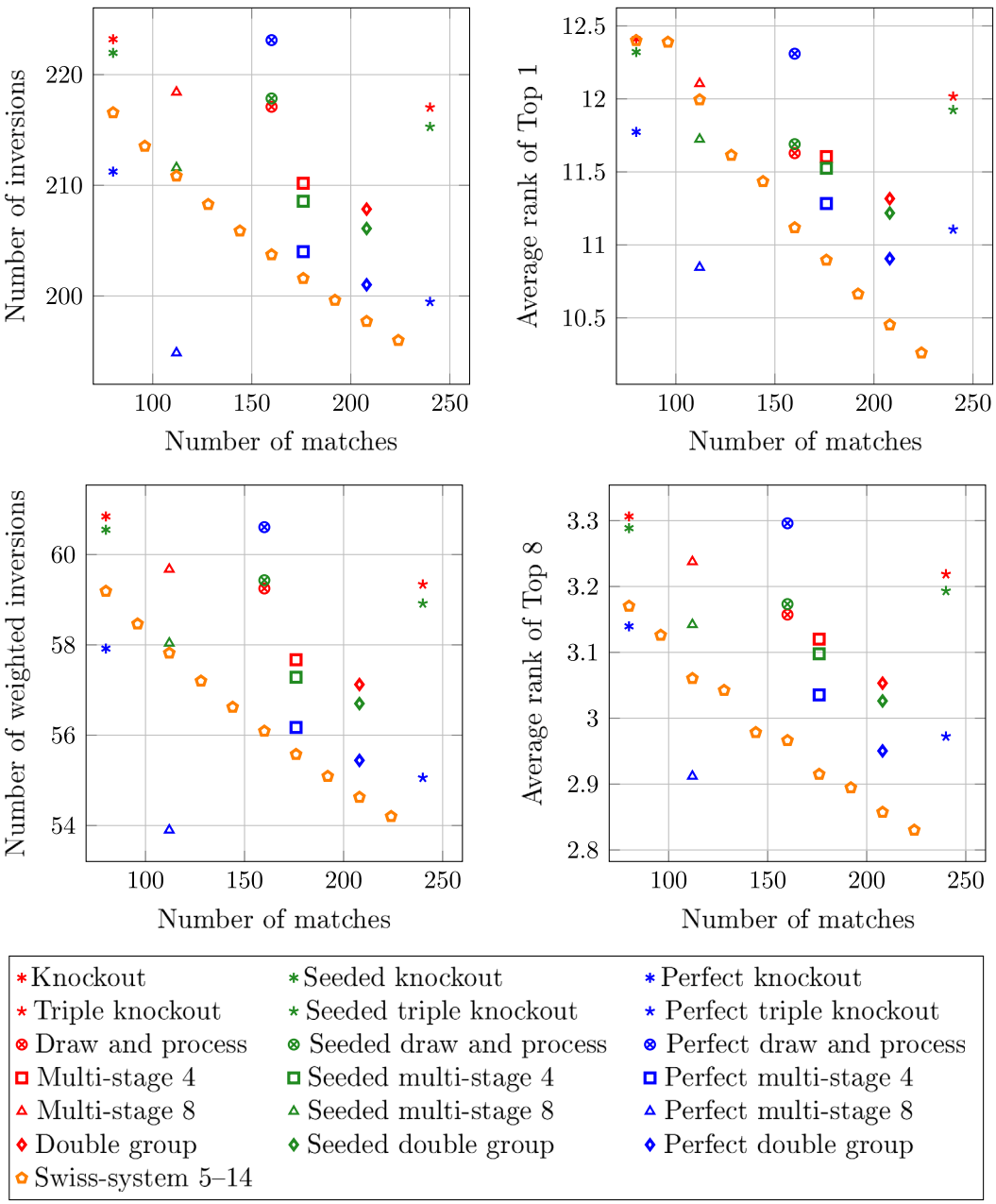


Fig. 8. Tournament metrics for the unseeded, realistically seeded (Seeded), and perfectly seeded (Perfect) designs.

4. Conclusions

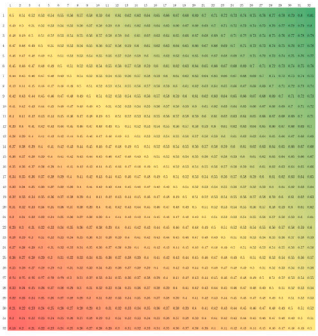
The Swiss-system has been shown to be an efficacious and reasonable alternative to traditional tournament formats. Its advantage is probably explained by taking directly the outcomes of previous matches into consideration. This feature is similar to the knockout format but the negative effects of an unexpected loss are mitigated.

Consequently, the Swiss-system is worth adopting in further sports. The results can be especially interesting for emerging esports, where the tournament formats are not yet solidified. Given the expanding consumer base of online sports activities, these questions will continue to be relevant for quite some time. The replacement of traditional sport formats is also not unprecedented, cf. the design of the most prestigious club competition in European soccer, the UEFA Champions League, where a similar format will be used from the 2024/25 season (UEFA, 2021).

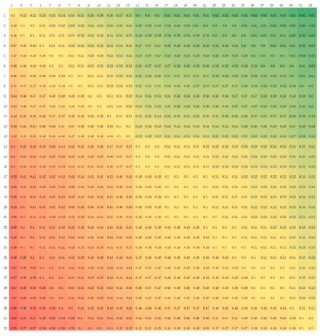
There remain several promising research directions. First, it remains an open question how ties affect the efficacy of the tournament designs considered. Second, the winning probabilities used in the paper are transitive but this assumption can be relaxed (Chen and Joachims, 2016). Third, each player plays the same number of matches in all our ranking mechanisms. Perhaps a similar accuracy can be achieved by removing certain clashes, especially if it is sufficient to rank the top players only. Fourth, as the Swiss-system turns out to be a competitive tournament format, connected issues such as the optimal ranking (Csató, 2013, 2017) and the details of the pairing algorithm (Biró et al., 2017; Führlich et al., 2021) may require further investigation. Finally, one may abstract from the traditional tournament formats and consider any sequence of pairwise comparisons. This direction is closely related to a research line in machine learning, where the problem is to retrieve the ranking of items from minimal number of noisy comparisons, see e.g. Ren et al. (2019).

Table A.1
Heatmap of the winning probabilities. The cell (i, j) shows the winning probability of player i against player j.

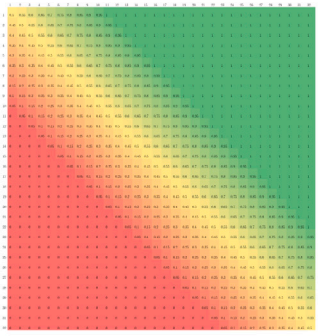
a `skill` $= 1$



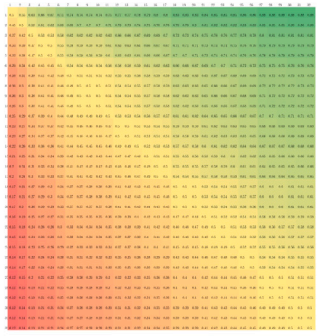
b Chess



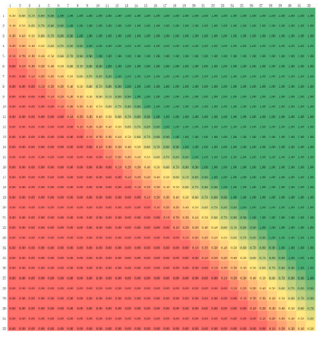
c `skill` $= 5$



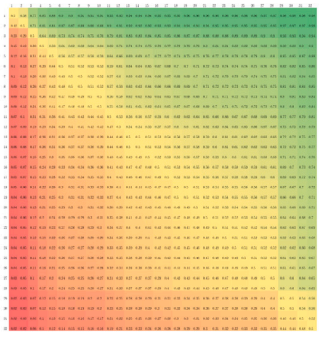
d Soccer



e `skill` $= 10$



f Tennis



CRediT authorship contribution statement

Balázs R. Sziklai: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Péter Biró:** Conceptualization, Methodology, Writing – review & editing. **László Csató:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgements

Six anonymous reviewers provided valuable comments and suggestions on earlier drafts.

The research was supported by the MTA Premium Postdoctoral Research Program grant PPD2019-9/2019, and by the Hungarian National Research, Development and Innovation Office, grant numbers K128573, K128611, and K138945.

Balázs R. Sziklai is the grantee of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the New National Excellence Program Bolyai+ scholarship of the Ministry for Innovation and Technology.

Péter Biró acknowledges the financial support by the Hungarian Academy of Sciences, Momentum Grant No. LP2021-2.

Appendix

See Table A.1.

References

Adler, I., Cao, Y., Karp, R., Peköz, E.A., Ross, S.M., 2017. Random knockout tournaments. Oper. Res. 65 (6), 1589–1596.

Ales, L., Cho, S.-H., Körpeoğlu, E., 2017. Optimal award scheme in innovation tournaments. Oper. Res. 65 (3), 693–702.

Annis, D.H., Wu, S.S., 2006. A comparison of potential playoff systems for NCAA IA football. Amer. Statist. 60 (2), 151–157.

Appleton, D.R., 1995. May the best man win? J. R. Statist. Soc.: Ser. D (Statist.) 44 (4), 529–538.

Arlegi, R., 2021. How can an elimination tournament favor a weaker player? Int. Trans. Oper. Res. http://dx.doi.org/10.1111/itor.12955, (in press).

Arlegi, R., Dimitrov, D., 2020. Fair elimination-type competitions. European J. Oper. Res. 287 (2), 528–535.

Atan, T., Çavdaroğlu, B., 2018. Minimization of rest mismatches in round robin tournaments. Comput. Oper. Res. 99, 78–89.

Bergantiños, G., Moreno-Ternero, J.D., 2020. Sharing the revenues from broadcasting sport events. Manage. Sci. 66 (6), 2417–2431.

Bimpikis, K., Ehsani, S., Mostagir, M., 2019. Designing dynamic contests. Oper. Res. 67 (2), 339–356.

Biró, P., Fleiner, T., Palincza, R.P., 2017. Designing chess pairing mechanisms. In: Frank, A., Recski, A., Wiener, G. (Eds.), Proceedings of the 10th Japanese-Hungarian Symposium on Discrete Mathematics and its Applications. Budapesti Műszaki s Gazdaságtudományi Egyetem, Budapest, pp. 77–86.

Blizzard Entertainment, 2020. Hearthstone tournament player handbook v2.5. 15 October. https://assets.blz-contentstack.com/v3/assets/bltc965041283bac56c/blt7180e70eed2edb7c/5f88cab321bd3d0cf67e06d6/Hearthstone_Tournament_Player_Handbook_v2.5.pdf.

Brown, J., Minor, D.B., 2014. Selecting the best? Spillover and shadows in elimination tournaments. Manage. Sci. 60 (12), 3087–3102.

Can, B., 2014. Weighted distances between preferences. J. Math. Econom. 51, 109–115.

Cea, S., Durán, G., Guajardo, M., Sauré, D., Siebert, J., Zamorano, G., 2020. An analytics approach to the FIFA ranking procedure and the world cup final draw. Ann. Oper. Res. 286 (1–2), 119–146.

Chater, M., Arrondel, L., Gayant, J.-P., Laslier, J.-F., 2021. Fixing match-fixing: Optimal schedules to promote competitiveness. European J. Oper. Res. 294 (2), 673–683.

Chen, R., Hwang, F.K., 1988. Stronger players win more balanced knockout tournaments. Graphs Combin. 4 (1), 95–99.

Chen, S., Joachims, T., 2016. Predicting matchups and preferences in context. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 775–784.

Croquet Association, 2021. Regulations for tournaments 2021. https://www.croquet.org.uk/?p=tournament/regulations.

Csató, L., 2013. Ranking by pairwise comparisons for swiss-system tournaments. CEJOR Cent. Eur. J. Oper. Res. 21 (4), 783–803.

Csató, L., 2017. On the ranking of a Swiss system chess team tournament. Ann. Oper. Res. 254 (1–2), 17–36.

Csató, L., 2020. The incentive (in)compatibility of group-based qualification systems. Int. J. Gen. Syst. 49 (4), 374–399.

Csató, L., 2021a. Tournament Design: How Operations Research Can Improve Sports Rules. In: Palgrave Pivots in Sports Economics, Palgrave Macmillan, Cham, Switzerland.

Csató, L., 2021b. A simulation comparison of tournament designs for the World Men's handball championships. Int. Trans. Oper. Res. 28 (5), 2377–2401.

Csató, L., 2022. Quantifying incentive (in)compatibility: A case study from sports. European J. Oper. Res. http://dx.doi.org/10.1016/j.ejor.2022.01.042, (in press).

Dagaev, D., Sonin, K., 2018. Winning by losing: Incentive incompatibility in multiple qualifiers. J. Sports Econ. 19 (8), 1122–1146.

Dagaev, D., Suzdaltsev, A., 2018. Competitive intensity and quality maximizing seedings in knock-out tournaments. J. Combin. Optim. 35 (1), 170–188.

David, H.A., 1959. Tournaments and paired comparisons. Biometrika 46 (1/2), 139–149.

Durán, G., Guajardo, M., Sauré, D., 2017. Scheduling the South American qualifiers to the 2018 FIFA world cup by integer programming. European J. Oper. Res. 262 (3), 1109–1115.

Edwards, C.T., 1998. Non-parametric procedure for knockout tournaments. J. Appl. Stat. 25 (3), 375–385.

Engist, O., Merkus, E., Schafmeister, F., 2021. The effect of seeding on tournament outcomes: Evidence from a regression-discontinuity design. J. Sports Econ. 22 (1), 115–136.

FIDE, 2020. Handbook. https://handbook.fide.com/.

Führlich, P., Cseh, A., Lenzner, P., 2021. Improving ranking quality and fairness in Swiss-system chess tournaments. Manuscript. arXiv:2112.10522.

Ginsburgh, V.A., Van Ours, J.C., 2003. Expert opinion and compensation: Evidence from a musical competition. Amer. Econ. Rev. 93 (1), 289–296.

Glenn, W.A., 1960. A comparison of the effectiveness of tournaments. Biometrika 47 (3/4), 253–262.

Glickman, M.E., 2008. BayesIan locally optimal design of knockout tournaments. J. Statist. Plann. Inference 138 (7), 2117–2127.

Goossens, D.R., Beliën, J., Spieksma, F.C.R., 2012. Comparing league formats with respect to match importance in Belgian football. Ann. Oper. Res. 194 (1), 223–240.

Goossens, D.R., Spieksma, F.C.R., 2012. Soccer schedules in europe: an overview. J. Sched. 15 (5), 641–651.

Groh, C., Moldovanu, B., Sela, A., Sunde, U., 2012. Optimal seedings in elimination tournaments. Econom. Theory 49 (1), 59–80.

Guyon, J., 2015. Rethinking the FIFA World Cup^{TM} final draw. J. Quant. Anal. Sports 11 (3), 169–182.

Guyon, J., 2018. What a fairer 24 team UEFA Euro could look like. J. Sports Anal. 4 (4), 297–317.

Guyon, J., 2020. Risk of collusion: Will groups of 3 ruin the FIFA world cup? J. Sports Anal. 6 (4), 259–279.

Harary, F., Moser, L., 1966. The theory of round robin tournaments. Amer. Math. Monthly 73 (3), 231–246.

Harris, C., Vickers, J., 1987. Racing with uncertainty. Rev. Econom. Stud. 54 (1), 1–21.

Hartigan, J.A., 1968. Inference from a knockout tournament. Ann. Math. Stat. 39 (2), 583–592.

Hennessy, J., Glickman, M., 2016. BayesIan optimal design of fixed knockout tournament brackets. J. Quant. Anal. Sports 12 (1), 1–15.

Horen, J., Riezman, R., 1985. Comparing draws for single elimination tournaments. Oper. Res. 33 (2), 249–262.

Hou, T., Zhang, W., 2021. Optimal two-stage elimination contests for crowdsourcing. Transp. Res. E 145, 102156.

Hwang, F.K., 1982. New concepts in seeding knockout tournaments. Amer. Math. Monthly 89 (4), 235–239.

Israel, R.B., 1981. Stronger players need not win more knockout tournaments. J. Amer. Statist. Assoc. 76 (376), 950–951.

Karpov, A., 2016. A new knockout tournament seeding method and its axiomatic justification. Oper. Res. Lett. 44 (6), 706–711.

Karpov, A., 2018. Generalized knockout tournament seedings. Int. J. Comput. Sci. Sport 17 (2), 113–127.

Kendall, G., Knust, S., Ribeiro, C.C., Urrutia, S., 2010. Scheduling in sports: An annotated bibliography. Comput. Oper. Res. 37 (1), 1–19.

Kendall, G., Lenten, L.J.A., 2017. When sports rules go awry. European J. Oper. Res. 257 (2), 377–394.

Klumpp, T., Polborn, M.K., 2006. Primaries and the new hampshire effect. J. Public Econ. 90 (6–7), 1073–1114.

Knuth, D.E., Lossers, O.P., 1987. A random knockout tournament. SIAM Rev. 29 (1), 127–129.

Koning, R.H., 2011. Home advantage in professional tennis. J. Sports Sci. 29 (1), 19–27.

Kräkel, M., 2014. Optimal seedings in elimination tournaments revisited. Econ. Theory Bull. 2 (1), 77–91.

Krumer, A., 2020. Testing the effect of kick-off time in the UEFA Europa League. Eur. Sport Manage. Q. 20 (2), 225–238.

Kulhanek, T., Ponomarenko, V., 2020. Surprises in knockout tournaments. Math. Mag. 93 (3), 193–199.

Laliena, P., López, F.J., 2019. Fair draws for group rounds in sport tournaments. Int. Trans. Oper. Res. 26 (2), 439–457.

Lasek, J., Gagolewski, M., 2018. The efficacy of league formats in ranking teams. Statist. Model. 18 (5–6), 411–435.

Lazear, E.P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. J. Polit. Econ. 89 (5), 841–864.

Marchand, E., 2002. On the comparison between standard and random knockout tournaments. J. R. Statist. Soc.: Ser. D (Statist.) 51 (2), 169–178.

McGarry, T., Schutz, R.W., 1997. Efficacy of traditional sport tournament structures. J. Oper. Res. Soc. 48 (1), 65–74.

Mendonça, D., Raghavachari, M., 2000. Comparing the efficacy of ranking methods for multiple round-robin tournaments. European J. Oper. Res. 123 (3), 593–605.

Orrison, A., Schotter, A., Weigelt, K., 2004. Multiperson tournaments: An experimental examination. Manage. Sci. 50 (2), 268–279.

Palacios-Huerta, I., Volij, O., 2009. Field centipedes. Amer. Econ. Rev. 99 (4), 1619–1635.

Pauly, M., 2014. Can strategizing in round-robin subtournaments be avoided? Soc. Choice Welf. 43 (1), 29–46.

Petróczy, D.G., Csató, L., 2021. Revenue allocation in formula one: A pairwise comparison approach. Int. J. Gen. Syst. 50 (3), 243–261.

Prendergast, C., 1999. The provision of incentives in firms. J. Econ. Lit. 37 (1), 7–63.

Preston, I., Szymanski, S., 2003. Cheating in contests. Oxf. Rev. Econ. Policy 19 (4), 612–624.

Prince, M., Cole Smith, J., Geunes, J., 2013. Designing fair 8- and 16-team knockout tournaments. IMA J. Manag. Math. 24 (3), 321–336.

Ren, W., Liu, J., Shroff, N.B., 2019. On sample complexity upper and lower bounds for exact ranking from noisy comparisons. Manuscript. arXiv:1909.03194.

Rosen, S., 1986. Prizes and incentives in elimination tournaments. Amer. Econ. Rev. 76 (4), 701–715.

Rubinstein, A., 1980. Ranking the participants in a tournament. SIAM J. Appl. Math. 38 (1), 108–111.

Ryvkin, D., 2010. The selection efficiency of tournaments. European J. Oper. Res. 206 (3), 667–675.

Ryvkin, D., Ortmann, A., 2008. The predictive power of three prominent tournament formats. Manage. Sci. 54 (3), 492–504.

Scarf, P.A., Yusof, M.M., 2011. A numerical study of tournament structure and seeding policy for the soccer world cup finals. Stat. Neerl. 65 (1), 43–57.

Scarf, P., Yusof, M.M., Bilbao, M., 2009. A numerical study of designs for sporting contests. European J. Oper. Res. 198 (1), 190–198.

Schwenk, A.J., 2000. What is the correct way to seed a knockout tournament? Amer. Math. Monthly 107 (2), 140–150.

Searls, D.T., 1963. On the probability of winning with different tournament procedures. J. Amer. Statist. Assoc. 58 (304), 1064–1081.

Szymanski, S., 2003. The economic design of sporting contests. J. Econ. Lit. 41 (4), 1137–1187.

Taylor, C.R., 1995. Digging for golden carrots: An analysis of research tournaments. Amer. Econ. Rev. 85 (4), 872–890.

UEFA, 2021. New format for Champions League post-2024: everything you need to know. 25 May. https://www.uefa.com/uefachampionsleague/news/0268-12157d69ce2d-9f011c70f6fa-1000--new-format-for-champions-league-post2024-everything-you-need-to-know/.

Van Bulck, D., Goossens, D., 2020. Handling fairness issues in time-relaxed tournaments with availability constraints. Comput. Oper. Res. 115, 104856.

Vong, A.I.K., 2017. Strategic manipulation in tournament games. Games Econom. Behav. 102, 562–567.

Vu, T., Shoham, Y., 2011. Fair seeding in knockout tournaments. ACM Trans. Intell. Syst. Technol. (TIST) 3 (1), 9.

Wright, M., 2014. OR analysis of sporting rules – A survey. European J. Oper. Res. 232 (1), 1–8.

Yücesan, E., 2013. An efficient ranking and selection approach to boost the effectiveness of innovation contests. IIE Trans. 45 (7), 751–762.