Journal of Quantitative Analysis in Sports

Volume 2, *Issue* 3 2006 *Article* 1

Which Ball is the Roundest? - A Suggested Tournament Stability Index

Torbjörn Lundh*

*Mathematical Sciences at Chalmers and Göteborg University, torbjorn.lundh@chalmers.se

Which Ball is the Roundest? - A Suggested Tournament Stability Index*

Torbjörn Lundh

Abstract

All sports have components of randomness that cause the “best” individual or team not to win every game. According to many spectators this uncertainty is part of the charm when following a competition or a match. Have different sports more or less of this unpredictability? We suggest here a general measure, a tournament stability index, together with its associated p-value which we denote the ”coin-tossing-index.” These indexes are aimed to quantify the randomness factor for different tournaments, and different sports. As an illustration we exemplify and discuss these measures for basketball, squash, and soccer. Some additional results will also be given on a few tournaments in ice-hockey, and handball. Furthermore, we discuss a couple of combinatorial optimization questions that turned up on the way.

KEYWORDS: tournament, ranking, combinatorial optimization, league standing effect, competitive balance, paired comparison, Slater’s i, Kendall’s u, nearest adjoining order

*The author wants to thank the two anonymous referees for many valuable and constructive comments and suggestions, both on the technical content and on the organization of the paper. Furthermore, the author is in dept to Sven-Erick Alm for encouraging discussions over the years and for the formula (7) derived together with Allan Gut. Last but not least, Andrew Plater of Cambridge University for helpful discussions concerning NP-completeness in general, and for informing the author of the second Bundesliga match-fixing scandal in particular.

1 Introduction

We have probably all heard sport commentators saying something like: "The ball is round and can go either way" or "That's the way the ball bounces."

How to quantify this unpredictability? The underlying idea we use is very simple: How often will a "better" team lose against a "weaker" opponent in a tournament?

Suppose that there is a ranking list[1] ρ of a group of n teams or individuals that play a tournament. Let a and b be two teams (or individuals) that play each other in game i in the tournament. Assume that a is ranked higher than b, i.e. $\rho(a) < \rho(b)$. We put a value, v_i, on this game i according to the following scheme:

$$v_i = \begin{cases} 1 & \text{if } a \text{ wins} \\ -1 & \text{if } b \text{ wins} \\ 0 & \text{if there is a draw.} \end{cases} \tag{1}$$

This evaluation is used in the so called Just Win Baby, "JWB", ranking system; see [26] for example.

We get the tournament index if we sum up all matches according to the scheme (1) above and divide the total by the number of matches played. That is, if a total of N decided[2] matches in the tournament are played, the tournament index, $\mathcal{T}$, is defined as

$$\mathcal{T}(\rho) = \frac{\sum_{i=1}^{k} v_i}{N}. \tag{2}$$

We have immediately that $-1 \leq \mathcal{T}(\rho) \leq 1$ and that $\mathcal{T}(\rho)$ is close to 1 if the ranking ρ is "correct" and there is not much randomness in the game. On the other hand, if there is much randomness, $\mathcal{T}(\rho)$ will be close to zero. Furthermore, if $\mathcal{T}(\rho)$ is close to -1, then $n - \rho + 1$ would be a good ranking, where n is the number of teams or players.

1.1 Related studies

A high degree of uncertainty regarding the outcome of a game is highly desirable to the owners of a league due to the economics of professional sports. This

[1] I.e. ρ is a bijection from the set of n teams to {1,2,...,n}.

[2] That is, the draws are not counted. In Sections 3.3 and 3.4 we illustrates how counting, or not counting, the draws effects the indexes.

quality was denoted the *league standing effect* by Neale in [15], and usually quantified using a measure called *competitive balance*, see for example [18] and [10]. We recall the definition and comment on the relations to our suggested index $\mathcal{T}$ in Section 2.9 below.

The question if the winner of the English Premier League is really the best team, is addressed in [28, Chapter 7], where among other things, a simulated random final league table is presented. We study other simulated Premier League tournaments in Section 3.3.

Paired comparison has long been a popular method for example in psychological studies, see for example [8]. An individual is given two options and have to choose one. For example, the test-person gets two glasses of wine and have to pick the tastiest of the two. Then this is repeated for all possible paired combinations among the wines to be ranked. Hence the method of paired comparison is closely related to sport tournaments' outcomes; see for example [6] and [1]. Of special interest to us is Kendall's u defined in [11], and Slater's i defined in [24]. These measures is recalled, discussed and compared to our index in sections 2.10 and 2.2.

After the first version (i.e. [13]) of this paper was submitted, a similar index as $\mathcal{T}$ was presented in [3] using an impressively large number of games of soccer, football, baseball, ice-hockey, and basketball; where the index was chosen as the frequency of "upsets", i.e. when a team defeats a higher ranked team, and where the ranking was picked as the current standing in the league, and thus updated after every game. Note that in the early stages of a tournament, the current standing might not so well reflect the real strengths of the teams. On the other hand, if the initial standing, ρ, obtained after the first round of games, was preserved throughout the whole series, then the upset index is $\frac{1-\mathcal{T}(\rho)}{2}$. In that sense, for "stable" tournaments, one could use this relation as an approximation of the upset index.

2 Suggested Measures of Tournament Stability

The index depends heavily on the ranking we choose, see for example the last paragraph in Section 3.1. To get around that problem, we use an after–ranking or quite simply, a result–list. Remember that we are interested in the stability of an already completed tournament, not to predict any future result, which is what rankings are usually supposed to do.

We pick an after–ranking based on the number of games won, and if that number is equal for two or more teams (or individuals), the internal meetings will decide which team is ranked higher. If teams have the same number of internal victories, then those teams will be randomly ranked[3]. If ρ_r is this result ranking, we denote $\mathcal{T}(\rho_r)$ by $\mathcal{T}_r$.

Using this after–ranking, we can expect a high tournament index; we cannot, however, always expect to get the highest possible index result by using this result–list ranking. That is, in some cases, there is an optimal ranking ρ_o such that $\mathcal{T}_o = \mathcal{T}(\rho_o) > \mathcal{T}_r$. We will come back to this peculiarity in section 2.1 below, but for now, let us concentrate on $\mathcal{T}_r$.

A problem with using the tournament's result–list as a ranking for studying the stability of the same tournament is that the index will be biased. For example, even if all games were decided by coin flipping, we would of course get a non-negative tournament index.

Suppose n teams meet every other team m times in a tournament, where all games were settled by coin-tossing. Let $M_r(n,m)$ denote the expected value of the index $\mathcal{T}_r$ of such a random tournament. In Table 2 in Appendix B, we give approximate values of $M_r(n,m)$ using Monte Carlo simulations.

To make up for the internal bias we introduce by choosing an after–ranking, we define a normalized result tournament index, $\widehat{\mathcal{T}_r}$, by a translation and rescaling of $\mathcal{T}_r$ in the following way,

$$\widehat{\mathcal{T}_r} = \frac{\mathcal{T}_r - E(\mathcal{T}_r)}{1 - E(\mathcal{T}_r)}, \tag{3}$$

where $\mathcal{T}_r$ is calculated as in (2) games played and using the result–ranking ρ_r. We note that $\widehat{\mathcal{T}_r} \leq 1$, and that the expected value of $\widehat{\mathcal{T}_r}$ would be zero, if the outcome of all games was decided randomly. We can use the index $\widehat{\mathcal{T}_r}$ as a measure of the stability of a tournament.

2.1 Optimal ranking

We can represent a tournament with a $n \times n$–matrix A with elements $a_{ij} \geq 0$ denoting the number of victories team i has against team j among the n teams in the tournament. Since the teams do not meet themselves, the diagonal will

[3] Note that many different rankings has been developed in order to take into consideration more involved structures, or incomplete tournaments etc, e.g. [6], [26], and [17].

be zero. Let us view the ranking ρ as a permutation of $(1, 2, 3, \ldots, n)$. Then we have that

$$\mathcal{T}(\rho) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n} sgn\big(\rho(j) - \rho(i)\big) a_{ij}, \tag{4}$$

where $N = \sum_i \sum_j a_{ij}$, i.e. the total number of decided games, and $sgn(\cdot)$ is the sign function. Then an optimal ranking is a ranking ρ_o that gives the largest index, i.e.

$$\mathcal{T}_o = \max_{\rho} \mathcal{T}(\rho) = \mathcal{T}(\rho_o). \tag{5}$$

We call the $\mathcal{T}_o$ the optimal tournament index. Note that even though the optimal ranking might not be unique, $\mathcal{T}_o$ is unique.

2.2 Slater's i

If all possible combinations, in a paired comparison sequence, are tested once, one would have a tournament like matrix, with only ones and zeroes; see [24]. The so called *nearest adjoining order* will be the ranking which gives the fewest number of inconsistencies in the matrix. This smallest number of inconsistencies is called Slater's i. Where inconsistence means an instance where test-person breaks the order by his choice. This ranking, which might not be unique, will also be an optimal ranking for $\mathcal{T}$ and we get the following relation between Slater's i and $\mathcal{T}_o$.

$$\mathcal{T}_o = 1 - \frac{2i}{N},$$

where N is the total number of games in the tournament. For example in the round-robin case where each of the n team meets each other m times and there are no draws, $N = \frac{(n-1)nm}{2}$.

We now give two examples where the usual result list, i.e. a team with more victories will be ranked higher than one with fewer wins, does not give the highest index. That is, examples where $\mathcal{T}_o > \mathcal{T}_r$.

2.3 Example 1 — three teams meeting each other three times.

Suppose such a tournament gives the following result matrix:

$$A = \begin{pmatrix} 0 & 1 & 3 \\ 2 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}.$$

The usual result list, or after–ranking, will be $(1, 2, 3)$. Using this ranking we get an index $\mathcal{T}_r = \frac{1}{9}$. But if we instead pick the ranking $(2, 1, 3)$ we get $\mathcal{T}_o = \frac{1}{3}$. We see that we can increase the index by switching places of teams that have almost the same number of total victories. The reason in this case is that team 2 has two wins and one losses against team 1.

In the following example, we limit ourselves to tournaments with just one match per pair. We then have to increase the number of teams to five in order to find an example where the result list will not give the optimal ranking.

2.4 Example 2 — five teams meeting each other only once.

Suppose the tournament matrix will be

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

then one, of three possible, result–list rankings will be $(4, 1, 2, 3, 5)$ giving the index $\mathcal{T}_r = \frac{2}{5}$. (The reason why there is more than one possible result is that team 1, 2, and 3 all have two wins each and one internal win among each other, i.e. 1 won over 2 which in turn won over 3 who won over 1.) However, the ranking $(4, 3, 1, 5, 2)$ gives a higher index $\mathcal{T}_o = \frac{3}{5}$. Note that team 5's only victory was against team 2.

2.5 How to find the optimal ranking?

Question 1 *Is the problem of finding the optimal ranking in (5) NP-complete?*

Due to the similarities to well known NP-complete problems such as *the (directed) optimal linear arrangement*, c.f. [9, p. 200]; *the quadratic assignment problem*, c.f. [22] and [9, p. 218], the author would be very surprised if the answer to that question would be no, even for the case where the teams just meet each other once. The discussion in [21] gives more arguments for this viewpoint. See also other related problems in the three volumes of [23].

2.6 Alway's algorithm

In [24] on p. 308, an algorithm for searching for an optimal order of tournament-like matrices with only zeroes and ones is described. This algorithm is effective, but not perfect as was shown in [20] using a 10×10 matrix.

Nevertheless, we have implemented Alway's algorithm with a slight generalization to tournament matrices including higher numbers than ones in order to get a first estimate of the optimal index; see Appendix A for a short description. In Table 3 in Appendix B, we have used this algorithm in a Monte Carlo simulation using 5000 random matrices to give approximations of the expectations, M_A, and variances of $\mathcal{T}_A$ for round-robin tournaments. Note that M_A is a lower estimation of M_o. There has been further generalizations and many algorithm constructions of this problem, see for example [25], [19], [14], [27], [7], and [12].

Since we normalize our index using the same algorithm for both the tournament matrix itself and to estimate the expected random index, one should not expect a too big discrepancy between normalized indexes of different ranking systems. As an illustration of this, see for example Table 1 and compare the different $\widehat{\mathcal{T}_r}$ and $\widehat{\mathcal{T}_A} \approx \widehat{\mathcal{T}_o}$ values.

2.7 The expected value of the optimal tournament index $\mathcal{T}_o$ for a random tournament

Intuitively, one might argue that $M_o(n, m)$ will decrease when the number of matches, m, increases since the difference between the artificial teams will be leveled out when there are more coin tosses. Similarly, we might expect $M_o(n, m)$ to decrease when the number of teams, n, increases, since it will be harder to find a clear ranking when more teams are involved. The result in Table 3 supports these arguments.

In the simple case where we just have two teams we can give a closed expression for the expected random tournament index.

$$M_o(2, m) = \frac{(m-1)!}{2^m} \sum_{i=0}^{m} \frac{|m-2i|}{i!(m-i)!}. \tag{6}$$

Note here that if the number of matches m is an even number, then $M_o(2, m+1) = M_o(2, m)$.

This formula (6) was later simplified by Sven-Erick Alm and Allan Gut of Uppsala University after a seminar there 2003, to the following form.

$$M_o(2, m) = \frac{1}{2^{m-1}} \mathrm{Bin}(m-1, [\frac{m-1}{2}]), \tag{7}$$

where $[\cdot]$ stands for the integer part.

Furthermore, we have only two possible outcomes of the optimal index value for simple round-robin tournaments with three teams and $M_o(3, 1) = 1\frac{3}{4} + \frac{1}{3}\frac{1}{4} = \frac{5}{6}$, with variance $\frac{1}{12}$. Similarly, for simple round-robin tournaments with four teams we have, as in the case above, only two possible outcomes for the index and $M_o(4, 1) = 1\frac{3}{8} + \frac{2}{3}\frac{5}{8} = \frac{19}{24}$, with variance $\frac{5}{192}$.

Question 2 *Is it possible to find closed expression for the expected optimal tournament index $M_o(n, m)$ for higher combinations of n and m?*

For higher number of teams and matches, one will get a distribution index value which seems to approach a normal distribution with mean and variance approximated in Table 3.

On the other hand, for small tournaments the distribution can not be estimated well using a normal distribution. For example in the case above with four teams and a simple round-robin, there is a high probability, i.e. $\frac{3}{8}$, to reach a situation with $\mathcal{T}_o = 1$. Hence a high index does not necessarily mean a very stable, i.e. non-stochastic, tournament for small tournaments. However, the *coin tossing index*, which is described just below, can be used to overcome such difficulties when comparing tournaments with different schemes and small sizes.

2.8 The Coin Tossing Index

We might want to use a p-value associated to our suggested index $\mathcal{T}_o$ in order to get an alternative viewpoint of what a specific index value indicates. We might also get a more structure independent measure, where we can better compare different tournaments which has different size, or constructions. For example, we might want to compare a season which consists of a sequence of knock-out cups, with a round-robin tournament with more teams. See for example Table 1. Such a p-value, which we denote the Coin Tossing Index, **CTI**, is the probability that a random tournament of the same size[4] and structure would give an equal or higher optimal tournament index $\mathcal{T}_o$. With a random tournament, we mean a tournament where all matches are decided randomly with equal weight, i.e. using a (well balanced) coin-toss. There is a drawback with this index for more stable tournaments since the **CTI** will then be so small that it is hard to estimate accurately using Monte Carlo simulations. However, one way to overcome such problems is to use a normal approximation of the index distribution.

2.9 Comparison with the measure of competitive balance

Let us study a tournament with n teams and let w_i be the winning frequency for team i, i.e. the number of victories divided by the number of games team i has played. The competitive balance is then measured using

$$\sigma_L = \sqrt{\frac{\sum_{i=1}^{n}(w_i - \frac{1}{2})^2}{n}}.$$

For more details on this see for example [10].

Both our tournament index and the competitive balance measures show how well ordered the tournament is, but there is no simple relation between them as the following four tournament matrices will illustrate. In the following examples, $\mathcal{T}$ stands for both $\mathcal{T}_r$ and $\mathcal{T}_o$.

$$\text{Let } A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

then $\sigma_L(A) \approx 0.3162$ and $\sigma_L(B) \approx 0.2236$. but $\mathcal{T}(A) = \mathcal{T}(B) = 0.8$. This illustrates the fact that the tournament index is only taken into consideration

[4] I.e. the same total number of games.

if a weaker team defeats a stronger team, not punishing the index more if the weaker team happens to be *much* weaker. See Section 2.12 below where we address this question using weights.

On the other hand if we let

$$C = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 2 & 0 & 0 \end{pmatrix}, \text{ and } D = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

then $\sigma_L(C) = \sigma_L(D) = 0$, but $\mathcal{T}(C) = \frac{1}{3}$ and $\mathcal{T}(D) = 0$. This example could be seen as an illustration how the tournament index is more sensitive to distributional changes, i.e. how complex the tournament matrix gets, in comparison to the competitive balance.

2.10 Comparison with Kendall's u function

Given a tournament matrix (a_{ij}) obtained by n teams meeting each other m times and where all games were decided, i.e. no draws. Let Σ be the sum of agreeing results between pairs of outcomes of games, i.e.

$$\Sigma = \sum_{i=1}^{n} \sum_{j=1}^{n} \binom{a_{ij}}{2}.$$

Then the Kendall's u function is defined as

$$u = \frac{2\Sigma}{\binom{n}{2}} - 1,$$

and is used to measure the amount of agreements in paired comparisons, see [11].

For 2×2-matrices we have a one-to-one correspondence between $\mathcal{T}$ $(= \mathcal{T}_r = \mathcal{T}_o)$ and the u functions. Let $m \geq 2$ and $n = 2$, then

$$u = (1 + c)\mathcal{T}^2 - c, \text{ where } c = \frac{m}{2\binom{m}{2}}.$$

However, for $n \geq 3$, there is no direct correspondence between $\mathcal{T}$ and u which the following example with $m = 2$ and $n = 3$ illustrates.

Let $A = \begin{pmatrix} 0 & 2 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 2 & 0 \end{pmatrix}$, and as above $C = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 2 & 0 & 0 \end{pmatrix}$

then $\mathcal{T}(A) = 1$, $u(A) = 1$, $\mathcal{T}(B) = \frac{1}{3}$, $u(B) = -\frac{1}{3}$, but $\mathcal{T}(C) = \frac{1}{3}$, and $u(C) = 1$.

2.11 Non round-robin tournaments

Let us finally mention one last question in this section. In the squash example we had a tournament over a year that was essentially composed by a series of (knockout) cups. This gave the consequence that the best player[5] also played the most games.

In a (pure knockout) cup with four teams the tournament matrix will look like this

$$A_{cup,4} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

listing the team in order of performance. It is straightforward to generalize this to any tournament size 2^k. In Section 3.2 below, we use 16 teams. In order to generate the right number of games, $A_{cup,16}$ is repeatedly added to a tournament matrix which randomly permutes the order of both the rows and columns for each cup added until the appropriate number of games is reached. In the case above with four teams three games will be added for each cup, and for the squash case below, 15 games are added for each cup of 16 teams.

2.12 A weighted ranking

The scheme (1) we have used so far to evaluate the outcome of a match is blunt in the sense that it punishes the score with -1 indifferently if for example the highest ranked team is beaten by the lowest ranked, as if it would have been beaten by the second highest ranked team.

[5]Peter Nicol

A way to get around this feature is to introduce a weighted ranking, x, in the following way. Let $x = (x_1, x_2, \ldots, x_n)$, where all $x_i \in [0, 1]$. Our new evaluation scheme of a given tournament $A = \{a_{ij}\}$, will be

$$\mathcal{W}(x) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j) a_{ij}, \tag{8}$$

where $N = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$. Compare this formulation with (4) and to Brown's set up in [5] where the weight is fixed to be the inverse rank, i.e. the team ranked as number one, gets weight n.

Now, let

$$\mathcal{W}_o = \max_{x \in [0,1]^n} \mathcal{W}(x). \tag{9}$$

We denote the optimal weighted ranking by x_o, i.e. $\mathcal{W}(x_o) = \mathcal{W}_o$, and where $x_o \in [0, 1]^n$.

Question 3 *Is the problem to find the optimal weighted ranking in (8) NP-complete?*

We could use $\mathcal{W}_o$ as an alternative stability index for tournaments, after it has been normalized as in (3) to $\widehat{\mathcal{W}_o}$.

Our naive strategy first to use Alway's algorithm to get a ranking. We then use this rankings and the lemma below to get candidates x with only zeroes and ones as their components. The ones are naturally set at the highest ranked positions. All these candidate vectors were then evaluated in Equation (8).

The following immediate result can be a tool in the investigation of Question 3 and to compute candidates for optimal weights allocations.

Proposition 4 *The optimal weighted ranking is in a (not necessary unique) corner in the unit hyper-cube, i.e.* $x_o \in \{0, 1\}^n$.

Proof. The partial derivative with respect to x_i of Equation (8) gives us immediately that the extremal value of $\mathcal{W}(x)$ has to be attained when x is in a corner in the unit hyper-cube.□

We will give an application of this weighted ranking in Section 3.3 below.

3 Applications

Let us now pick a few real world examples as illustrations to the above suggested indexes $\widehat{\mathcal{T}}_r$, $\widehat{\mathcal{T}}_o$ and **CTI**.

3.1 NBA 1995–1996

Let us start with basketball and the NBA season 1995–1996. 29 teams played 82 games each except the play-off teams who played up to 103 games in total (Seattle Super Sonics).

All teams met each other either two or four times before the play-off. In total there were 1189 games played. Using this data, we find $\mathcal{T}_r = 0.41$. In order to compute the normalized index, we need the expected value of an analogue tournament where all matches were decided randomly. Furthermore, using Alway's algorithm we get $\mathcal{T}_o \approx \mathcal{T}_A = 0.445$

This was done by randomly generating tournament matrices where all games met each other at least twice, and some four times. Doing this 10000 times we get estimates for $E(\mathcal{T}_r) \approx 0.094$ and $E(\mathcal{T}_A) \approx 0.19$. This gives us $\widehat{\mathcal{T}_r} \approx 0.34$ and $\widehat{\mathcal{T}_o} \approx \widehat{\mathcal{T}_A} \approx 0.31$.

As a comparison, consider a tournament where all 29 teams met exactly three times each. Such a tournament would give a total of 1218 games which can be compared with 1189. We can then use Table 2 where $M_r(29,3) \approx 0.14$. and we can estimate

$$\widehat{\mathcal{T}_r} \approx \frac{0.41 - 0.14}{1 - 0.14} \approx 0.31.$$

For further comparisons, let us also see what happens if we pick rankings ahead of the actual season. We look at two such examples. In those cases we do not normalize. With the ranking ρ based on the previous season, taking into consideration the actual points difference in each game, we get a tournament index of $\mathcal{T}(\rho) = 0.31$ which is very close to our $\widehat{\mathcal{T}_r}$. But if we instead choose a different ranking method which weights the different games according to the strength of the opponent (based on past meetings), we get instead $\mathcal{T}(\rho) = 0.073$, see [26] for more details on these and related rankings. Hence we see that the choice of ranking is essential.

3.2 Squash

Let us now exemplify the tournament index for an individual sport, namely squash, and more specifically the professional cups which are played around the world. The professional squash association, PSA, produces rankings of the players, see [2]. We pick the twenty highest ranked players from the list of 1st January 2002 and follow their results during the year 2001.

We record each game whenever two players from the list meet making a result matrix this way. In total we recorded 153 games this way. Using the result-ranking (which differs slightly from the PSA January 2002 ranking) we get $\mathcal{T} \approx 0.71$. Normalizing this, we find that $\widehat{\mathcal{T}}_r \approx (0.712-0.286)/(1-0.286) \approx 0.60$, where we used the normalization factor 0.286 taken from a Monte Carlo method of accumulated simulated cups of size 16 ($= 2^4 < 20 < 2^5$), repeated until we got 150 games in total. If we instead use Alway's algorithm we get $\widehat{\mathcal{T}}_o \approx \widehat{\mathcal{T}}_A \approx \frac{0.7255-0.387}{1-0.387} \approx 0.55$.

Alternatively, we can use Table 2 to see that in a tournament with 20 players where each one meets once, we get a total of 190 games and $M_r(20,1) = 0.34$, which gives a normalized index of 0.56 instead. In comparison, we see that if we use the ranking from January 2001 and follow the 20 highest ranked players during 2001, we get $\mathcal{T} = 0.26$ (which we do not normalize, since it is based on past information).

Comparing with the NBA example, where we got $\widehat{\mathcal{T}}_o \approx 0.34$, it seems that the professional top-squash during 2001 was more "stable" than the NBA season 1995-1996.

3.3 Premier League 2000–2001

Let us now turn to soccer, since we take a look at the English Premier League results during the season 2000–2001. We collect our data from [4]. Here, there were 20 teams playing each other 2 times each. That gives us a total of 380 games. A great part of them, 109, ended with a draw. Using our result-ranking, we get $\mathcal{T} = 0.34$ and $\widehat{\mathcal{T}}_r = 0.12$. Recall from (2) that N is the number of decided games, hence $N = 380 - 109$. If we instead divide by the total number of all games played, we would get $\mathcal{T}' = 0.24$, and from Table 2 we get $M_r(20,2) = 0.207$. Hence

$$\widehat{\mathcal{T}}' \approx \frac{0.24 - 0.207}{1 - 0.207} \approx 0.048.$$

We see there is a significant difference between these different approaches. In [13, Table 3] we used the practice to divide by the total number of all games, including the draws.

What does $\widehat{\mathcal{T}}_r = 0.12$ really mean? How close is this to complete randomness? The coin tossing index, $\mathbf{CTI} = 0.27$, gives a complementary view on this in the following sense. First, make a random tournament similar to the

Premier League 2000–2001 by letting 20 teams meeting each other twice and decide the winner by tossing a coin. After that, remove, at random places, 109 of these results, simulating the draws. Then the probability that this resulting random tournament will have a higher index $\mathcal{T}_A$ than the index for the real tournament is about 0.27. So in about one out of four random tournaments one would get a more structured tournament than this Premier League season.

If we do the same for the NBA and squash examples above we get **CTI**:s far less than 0.0001.

This result indicates that professional soccer is much more random than both basketball and squash. At least for these three tournaments studied.

What about the optimal weighted ranking? We can use Proposition 4 and propose a candidate for x_o with

$$x = (1,1,1,1,1,1,1,1,0,1,1,0,0,0,0,0,0,0,0,0) \quad \text{giving us } \mathcal{W}_o \geq 0.2546,$$

where the order is taken as the result ranking. We can then approximate $\widehat{\mathcal{W}_o} \approx 0.22$.

3.4 Bundesliga

Let us compare the Premier League result with another European professional soccer tournament, the German Bundesliga. We pick up the data from [4] and treat it in a similar way as above. This give us $\widehat{\mathcal{T}_o} \approx 0.044$ and $\widehat{\mathcal{T}_r} \approx 0.065$. If we count every game as above we would end up with $\mathcal{T}' = 0.24$ and from the Table 2, we have $M(18,2) = 0.22$, which gives us a normalized tournament index $\widehat{\mathcal{T}}' \approx 0.027$. That is even less than its English version! If we make coin tossing tournaments of the same size, we will in about a third of the random trials get a higher index.

4 Discussion – the nature of different sports

Our few tournament results listed in Table 1, might indicate that for example squash and basketball seem to be more stable sports than soccer, in the sense that the “better” player or team more often wins, compared to soccer, at least on a professional level. However, more tournament results need of course to be studied before one could make a more solid statement on this. As mentioned in the introduction, there has just recently been a related study carried out on

Sport	**Country**	**Season**	**Tournament**	$\widehat{\mathcal{T}}_r$	$\widehat{\mathcal{T}}_A$	**CTI**
Soccer	England	00/01	Premier L.	0.12	0.050	0.27
Soccer	England	02/03	Premier L.	0.21	0.12	0.056
Soccer	England	03/04	Premier L.	0.22	0.18	0.003
Soccer	Germany	00/01	Bundesliga	0.065	0.044	0.31
Soccer	Germany	02/03	Bundesliga	0.041	0.048	0.27
Soccer	Germany	03/04	Bundesliga	0.19	0.14	0.022
Soccer	Germany	02/03	2 Bundesliga	0.14	0.17	0.015
Soccer	Germany	03/04	2 Bundesliga	-0.053	-0.12	0.95
Soccer	France	03/04	Division 1	0.20	0.13	0.021
Soccer	Spain	03/04	Division 1	0.20	0.16	0.006
Soccer	Italy	03/04	Seria A	0.36	0.32	<0.001
Ice-hockey	Switzerland	03/04	National L.	0.22	0.20	<0.001
Ice-hockey	Germany	03/04	DEL	0.079	0.022	0.35
Handball	Germany	03/04	Bundesliga	0.47	·	<0.0001
Basketball	USA	96/97	NBA	0.34	0.31	<0.0001
Squash	Intern.	01	PSA	0.60	0.55	<0.0001

Table 1: This is a comparison between different tournaments from different sports, countries, and years. We list approximations of the three suggested indexes, the normalized touring index with respect to the result list $\widehat{\mathcal{T}}_r$, the normalized touring index with respect to the optimal ranking $\widehat{\mathcal{T}}_o$, and the *Coin Tossing Index*, **CTI**, which is the expected probability for a random tournament to have a higher $\widehat{\mathcal{T}}_o$ than obtain in the tournament in question. The estimates are based on Monte-Carlo simulations using 10 000 matrixes.

a very high number of matches in [3], verifying that soccer is indeed in general more uncertain than football, baseball, ice-hockey, and basketball.

What could then be the causes? One obvious reason is that in soccer there usually are not so many chances to score, hence a single fluke play might have a greater impact to the outcome of the game.

Another reason might be that the level of the top soccer players is extremely high and even. There are very few natural talents in that sport that are not taken care of at an early stage. Many children play with a soccer-ball in some form all over the world, but not that many have ever seen a squash ball.

By measuring more basketball, squash, soccer, and other tournaments, one would ask if there might be some universal numbers of the randomness for the different sports. How do professional series differ from amateur tournaments? Maybe there is an interval where the tournament index should lie to become an attractive public sport such as the league standing effect described in [15]? Maybe this interval differs from person to person? How often do we want "David to defeat Goliath?" In some sense you can view the rate of randomness in a tournament as a measure of competitive and exciting the series is, but that cannot be the whole answer, otherwise coin-tossing would be the most spectacular sports of all. There are of course other criterions how we can compare different sports, see for example Chapter 1 in [16].

Finally we would like to mention the outlier in Table 1 where it is indicated that about 95% of the random tournaments will end up with a higher optimal index than the second Bundesliga 2003-2004. (Note that both the normalized indexes where negative for this tournament.) We where quite concerned about this strange result until we learnt that there were some manipulations of that tournaments in forms of alleged match-fixing supposedly involving players, and a referee who was sentenced to jail. This indicates that our indexes might even be useful for monitoring sound tournaments.

References

[1] D.H. Annis and B.A. Craig *Hybrid Paired Comparison Analysis, with Applications to the Ranking of College Football Teams*, Journal of Quantitative Analysis in Sports, Vol. 1, No. 1, 2005.

[2] R. Beck et al. *SquashTalk*, Cup results and ranking in squash, www.squashtalk.com

[3] E. Ben-Naim, F. Vazquez and S. Redner, *What is the most competitive sport?* preprint, http://arxiv.org/abs/physics/0512143, December 2005.

[4] Professional League, www.ligaverwaltung.de

[5] E.L. Crow *Ranking paired contestants,* Communs. Statist.-Simula., 19, pp. 749–769 1990.

[6] H.A. David, *The Method of Paired Comparisons*, 2nd edn. Griffin, London, 1988.

[7] R. Delver and H. Monsuur, *Echelons in Incomplete Relations*, Theory and Decision, 44, pp. 279–292, 1998.

[8] G.T. Fechner, *Elemente der Psychophysik*, Breitkopf und Härtel, Leipzig, 1860. (English translation: *Elements of Psychophysics*, Holt, Rinehart & Winston, New York, 1965.)

[9] M.R. Garey and D.S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W.H. Freeman and company, San Fransisco, 1979.

[10] B.R. Humphreys *Alternative Measures of Competative Balance in Sports Leagues*, Journal of Sports Economics, Vol. 3 No. 2, pp. 133-148, May 2002.

[11] M.G. Kendall and B. Babington Smith *On the Method of Paired Comparasions*, Biometrica, Vol. 31, No. 3/4., pp. 324-345, 1940.

[12] L. Klukowski, *The nearest adjoining order method for pairwise comparisons in the form of difference of ranks*, Annals of Operations Research, 97, pp. 357–378, 2000.

[13] T. Lundh, *Which ball is the roundest? - a suggested tournament stability index*, Blå serien preprint, ISSN 1652-9715, 2005:28, www.math.chalmers.se/Math/Research/Preprints/2005/28.pdf, July 2005.

[14] A. Maas, T.G.G. Bezembinder, and P.P. Wakker, *On Solving Intransitivities in Repeated Pairwise Choices*, Mathematical Social Sciences 29, pp. 83-101, 1995.

[15] W.C. Neale *The peculiar economics of professional sports*, Quarterly Journal of Economics, 78, pp. 1-14, 1964.

[16] D. Oliver, em Basketball on paper, Brassey's Inc., 2004.

[17] J. Park and M.E.J. Newman, *A network-based ranking system for US college football,* Journal of Statistical Mechanics: Theory and Experiment, October 2005.

[18] J. Quirk, and R.D. Fort *Pay dirt: The business of professional team sports,* Princeton University Press, 1997.

[19] R.H. Ranyard, *An Algorithm for Maximul Likelihood Ranking and Slater's i from Paired Comparisons,* Br. J. math. statist. Pshychol., 29, pp. 242-248, 1976.

[20] R. Remage and W.A. Thompson *Maximum-likelihood paired comparison rankings*, Biometrika 53, pp. 143–149, 1966.

[21] J.M. Roberts, Jr., *Modeling hierarchy: Transitivity and the linear ordering problem*, J. Math. Sociol. 16 pp. 77–87, 1990.

[22] S. Sahni and T. Gonzalez, *P-Complete Approximation Problems*, Journal of the Association for Computing Machinery, Vol. 23, No. 3, pp. 555-565, July 1976.

[23] A. Schrijver, *Combiatorial Optimization*, Algorithms and Combinatorics 24, Springer-Verlag, 2003.

[24] P. Slater *Inconsistencies in a schedule of paried comparasions*, Biometrika, Vol. 48, 3 and 4, pp. 303–312, 1961.

[25] A.F.M. Smith and C.D. Payne, *An algorithm for determining Slater's i and all nearest adjoining orders,* Br. J. math. statist. Pshychol., 27, pp. 49–52, 1974.

[26] S.P. Sorensen, *An overview of some methods for ranking sports teams,* preprint `www.phys.utk.edu/sorensen/ranking/`, 1999.

[27] H. de Vries *Finding a dominance order most consistent with a linear hierarchy: a new procedure and review*, Anim. Behav., 55, pp. 827–843, 1998.

[28] J. Wesson, *The Science of Soccer*, Institute of Physics Publishing, Bristol and Philadelphia, 2002.

A Alway's algorithm

Let us give a brief description of this algorithm. Firstly, order the rows in the tournament matrix according to the result-ranking. Then examine the first line from the diagonal element and forward and count the numbers of wins and losses, i.e. $a_{1,j} - a_{j,1}$. If at some point, say at column j, the the accumulated losses are more than the accumulated wins, then transform the matrix by placing the first row at row j and column 1 to column j. Then start again from the new first row. This procedure is then repeated for all rows. When one reach the final row without any changes, one follows an analogous scheme for the columns, again starting at the diagonal element, but now going upward along that column. You will eventually get a good candidate for the optimal ranking. See more details in [24, p. 308].

B Estimates of $M_r(n,m)$ and $M_A(n,m)$

We give here two tables, with Monte Carlo estimates of the expectations M_r and M_A together with their variances, which might be useful if the reader wants to investigate tournaments of their own.

	m				
n	1	2	3	4	5
2	1.000; 0.000	0.501; 0.250	0.498; 0.83	0.372; 0.11	0.377; 0.056
3	0.842; 0.085	0.473; 0.077	0.417; 0.047	0.348; 0.037	0.316; 0.028
4	0.791; 0.026	0.458; 0.036	0.381; 0.021	0.314; 0.018	0.286; 0.014
5	0.713; 0.032	0.409; 0.023	0.343; 0.014	0.283; 0.011	0.259; .0088
6	0.642; 0.024	0.381; 0.016	0.318; .0098	0.262; .0077	0.236; .0062
7	0.596; 0.019	0.352; 0.012	0.294; .0073	0.243; .0060	0.219; .0045
8	0.552; 0.015	0.331; .0090	0.276; .0052	0.225; .0045	0.206; .0036
9	0.521; 0.012	0.310; .0071	0.261; .0043	0.215; .0036	0.193; .0027
10	0.489; .0092	0.295; .0057	0.246; .0035	0.204; .0028	0.183; .0022
11	0.467; .0078	0.281; .0048	0.234; .0029	0.194; .0024	0.175; .0018
12	0.445; .0068	0.270; .0041	0.225; .0025	0.184; .0021	0.169; .0016
13	0.426; .0059	0.259; .0036	0.216; .0021	0.178; .0018	0.161; .0013
14	0.409; .0048	0.250; .0031	0.206; .0019	0.171; .0015	0.156; .0011
15	0.393; .0046	0.240; .0027	0.201; .0017	0.166; .0013	0.149; .0010
16	0.380; .0040	0.233; .0023	0.195; .0014	0.160; .0011	0.145; .0009
17	0.368; .0036	0.224; .0020	0.187; .0012	0.155; .0010	0.141; .0008
18	0.356; .0031	0.218; .0019	0.183; .0011	0.151; .0009	0.137; .0007
19	0.345; .0028	0.211; .0017	0.178; .0010	0.147; .0008	0.133; .0006
20	0.336; .0026	0.207; .0015	0.173; .0009	0.144; .0007	0.130; .0006
21	0.327; .0023	0.201; .0014	0.169; .0008	0.140; .0007	0.127; .0005
22	0.319; .0021	0.198; .0013	0.166; .0008	0.136; .0006	0.123; .0005
23	0.310; .0021	0.193; .0012	0.161; .0007	0.133; .0006	0.121; .0004
24	0.305; .0018	0.187; .0011	0.158; .0007	0.130; .0005	0.118; .0004
25	0.298; .0017	0.185; .0010	0.154; .0006	0.127; .0005	0.116; .0004
26	0.290; .0016	0.180; .0009	0.151; .0006	0.126; .0004	0.113; .0003
27	0.286; .0014	0.176; .0008	0.149; .0005	0.122; .0004	0.111; .0003
28	0.280; .0014	0.173; .0008	0.145; .0005	0.120; .0004	0.109; .0003
29	0.274; .0013	0.169; .0007	0.143; .0005	0.118; .0003	0.107; .0003
30	0.269; .0012	0.167; .0007	0.141; .0004	0.116; .0003	0.106; .0003

Table 2: Approximations of expected tournament indexes for completely random games, $M_r(n,m)$, together with estimates of their variances, where the third digits should only be viewed as an indication. To illustrate this, note that Equations (6) or (7) gives the exact values for the $M_r(2,m)$–values on the first line which then really should read as 1.000 0.500 0.500 0.375 0.375 0.3125. We have used 5000 random matrixes in the Monte Carlo simulation for each pair n,m.

	m									
n	1		2		3		4		5	
2	1.000;	0.000	0.510;	0.250	0.505;	0.085	0.376;	0.110	0.376;	0.060
3	0.828;	0.085	0.474;	0.075	0.445;	0.037	0.354;	0.034	0.342;	0.022
4	0.789;	0.026	0.458;	0.035	0.413;	0.017	0.336;	0.016	0.316;	0.011
5	0.730;	0.024	0.443;	0.020	0.386;	0.011	0.317;	.0090	0.297;	.0069
6	0.685;	0.016	0.420;	0.013	0.368;	.0073	0.304;	.0059	0.279;	.0047
7	0.648;	0.013	0.401;	.0091	0.349;	.0053	0.289;	.0044	0.268;	.0034
8	0.618;	0.010	0.383;	.0068	0.334;	.0042	0.277;	.0034	0.255;	.0025
9	0.591;	.0081	0.371;	.0053	0.320;	.0033	0.268;	.0026	0.246;	.0021
10	0.567;	.0067	0.357;	.0044	0.310;	.0025	0.258;	.0021	0.237;	.0017
11	0.547;	.0055	0.347;	.0035	0.299;	.0022	0.249;	.0017	0.228;	.0014
12	0.530;	.0048	0.338;	.0029	0.289;	.0018	0.242;	.0014	0.222;	.0011
13	0.513;	.0043	0.328;	.0025	0.281;	.0016	0.235;	.0012	0.215;	.0010
14	0.496;	.0038	0.320;	.0022	0.274;	.0013	0.229;	.0010	0.210;	.0008
15	0.482;	.0032	0.312;	.0019	0.266;	.0012	0.224;	.0009	0.204;	.0007
16	0.470;	.0029	0.304;	.0016	0.260;	.0011	0.219;	.0008	0.199;	.0006
17	0.459;	.0025	0.298;	.0014	0.253;	.0010	0.213;	.0007	0.194;	.0006
18	0.449;	.0023	0.292;	.0012	0.248;	.0008	0.209;	.0006	0.191;	.0005
19	0.439;	.0020	0.286;	.0012	0.243;	.0007	0.205;	.0006	0.187;	.0004
20	0.429;	.0019	0.281;	.0010	0.238;	.0007	0.202;	.0005	0.183;	.0004
21	0.421;	.0017	0.276;	.0009	0.234;	.0006	0.198;	.0005	0.180;	.0004
22	0.415;	.0015	0.272;	.0008	0.229;	.0005	0.194;	.0004	0.177;	.0003
23	0.406;	.0014	0.268;	.0008	0.226;	.0005	0.191;	.0004	0.174;	.0003
24	0.399;	.0013	0.263;	.0007	0.222;	.0004	0.188;	.0003	0.171;	.0003
25	0.392;	.0012	0.260;	.0006	0.218;	.0004	0.186;	.0003	0.168;	.0003
26	0.386;	.0011	0.256;	.0006	0.215;	.0004	0.183;	.0003	0.166;	.0002
27	0.380;	.0011	0.252;	.0005	0.212;	.0004	0.180;	.0003	0.163;	.0002
28	0.375;	.0010	0.249;	.0005	0.209;	.0003	0.178;	.0003	0.161;	.0002
29	0.370;	.0009	0.245;	.0005	0.206;	.0003	0.176;	.0002	0.159;	.0002
30	0.364;	.0008	0.242;	.0004	0.203;	.0003	0.173;	.0002	0.156;	.0002

Table 3: Estimations of expected optimized tournament indexes for random games, $M_A(n,m)$, together with estimates of their variances. Note that $M_A(n,m)$ is a lower estimate of $M_o(n,m)$. We have here used 5000 random matrixes in the Monte Carlo simulation and Alway's algorithm, see [24, p. 308] for each pair n,m, where as usual m stands for the number of games each team play each other team, and n for the number of teams.