A Monte-Carlo analysis of competitive balance and reliability across tournament structures

Vishnu V. Nittoor
nitvishn@gmail.com

The International School Bangalore

September 2020

Abstract

This paper investigates the effect of increasing competitive balance on the reliability of tournament rankings. Reliability of rankings, a previously qualitative property, is quantified in this paper by the closeness between ground truth rankings and the rankings of teams at the end of a tournament. Three metrics are used to measure this closeness: Spearman's rank correlation coefficient, Kendall's tau, and a relatively unused algorithm in the field of ranking: Levenshtein distance. Three tournament structures are simulated: round-robin, random pairings, and the Swiss system. The tournaments are simulated across multiple trials and over a varying number of games. It is found that the rate of growth of reliability of a tournament structure falls as the number of games increases. It is also found that there is a positive relationship between competitive imbalance and reliability. The marginal benefit of increasing competitive imbalance falls as it is increased. Unexpectedly, in comparison to random pairings and Swiss pairings, the round-robin tournament structure is seen to achieve the highest reliability score across all metrics and number of games played. The difference in reliability between the tournament structures increases as competitive imbalance is increased. The further work suggested includes investigation of tournament outcome uncertainty in conjunction with reliability and competitive balance, a closer study into Levenshtein distance as a useful algorithm to quantify closeness between two rankings, and an inquiry into the specific factors that bottleneck reliability while the number of games played in a tournament increases.

1 Introduction

What makes the results from a tournament reliable? When the rankings at the end of a tournament are announced, and a winner is declared, how confident can we be that the rankings accurately represent the abilities of the teams? This question has long lingered in the minds of tournament organisers, competitors, and viewers. Although this question is rather intuitive to ask and begin to think about, quantifying reliability and the factors that affect it has remained relatively obscure. This paper aims to tap into that intuition and quantify reliability while numerically analysing two factors that affect it: competitive balance and tournament structures.

Reliability, in this paper, refers to how close the rankings obtained from the tournament are to the **true rankings** of the teams. It may not be clear how to investigate reliability in normal sports contexts as there exist wildly differing opinions on what the true rankings of the teams are in the first place, and so the relationship between the tournament rankings and the true rankings of teams is usually viewed as a grey area.

This paper, however, endeavours to tackle this problem by performing a Monte-Carlo simulation in which the true rankings of the teams are already known. In other words, outcomes of tournaments are generated from the inherent ground truths that have been set beforehand, allowing acute analysis of outcomes in the context of the true rankings from which they are probabilistically derived. This paper introduces a measure of reliability of rankings.

Competitive balance is defined by the level of disparity in competitors' abilities. If competitors in a tournament are of roughly equal strengths, the competitive balance is said to be high. On the other hand, if the competitors differ drastically in terms of their in-game strengths, a high competitive imbalance is said to exist.

Another measure related to reliability is tournament outcome uncertainty. Tournament outcome uncertainty is defined as the level of uncertainty in the tournament results; how much are the rankings liable to change if the tournament was conducted once more? Previous investigations by Scarf et al. 2009 [10] and Kringstad et al. 2004 [7] of competitive balance and tournament outcome uncertainty have suggested that increasing the level of competitive balance in a tournament increases the tournament outcome uncertainty. This paper is consistent with this result.

Furthermore, this paper also establishes that as competitive balance decreases, reliability increases. This paper details the nuances of this relationship while comparing the differences in tournament structures in achieving reliability with regards to the number of games played. It also presents a numerical perspective on how reliability varies across tournament structures over a range of sets of teams with different competitive balances.

1.1 The Overwatch League

The paper is based on performing a Monte-Carlo simulation of multiple tournament structures using data from the Overwatch League, a major esports league for the popular online video game *Overwatch*. This league consists of 20 teams and plays games over four stages, each lasting five weeks. At the end of each stage, the top eight teams engage in a playoff to determine a stage winner. At the end of the four stages, the top teams compete in a series of further playoffs which lead to the grand finals in which an overall winner is declared.

Data from the Overwatch League is used to simulate tournaments at a very low level of competitive imbalance. For higher levels of competitive imbalance, ratings are sampled from a Gaussian distribution with progressively larger standard deviations. More information about this is presented in section 4.2.

2 Methodology

This paper performs a Monte-Carlo simulation of different sets of teams in different tournament structures. The teams that provide the basic set of ratings are taken from the Overwatch League, and further sets of teams which have more competitive imbalance are generated from this data. Three tournament structures are used in this paper to investigate reliability and competitive balance.

1. Round robin
2. Random pairings of teams
3. The Swiss system

The different sets of teams will have different competitive balances or different levels of spread in the abilities of teams. Within each tournament structure, the effect of increasing the number of games played on reliability is investigated. The Elo rating system is used to set ground truth ratings for each set of teams. At the end of the tournament, they are ranked according to their win percentages. The following three metrics are used to measure reliability.

1. Spearman's Rank Correlation Coefficient
2. Kendall's Tau
3. Levenshtein Distance

These metrics are used to compare rankings between each tournament outcome and the ground truth to measure reliability.

3 Tournament structures

This paper investigates how competitive balance affects reliability within three different tournament structures. It also evaluates the merits of each tournament structure against each other in terms of multiple metrics.

3.1 The Round Robin Format

After the traditional knockout tournament system, one immensely popular tournament structure is the Round Robin format. In this format, every single team plays every other team a fixed number of times. It is the most game-intensive format observed in this paper: the number of games between teams is the highest here.

In this paper, the Round Robin format is implemented in a straightforward way with one modification: the number of times the teams are made to play each other is a variable. In the simulation that follows, the effect of increasing the number of rounds in the tournament on reliability and outcome uncertainty is investigated.

In a tournament system with n teams, there are exactly ${}^{n}C_2$ unique pairings between teams. Since each team plays every other team *once* during a single round of the tournament, it is observed that there are exactly ${}^{n}C_2$ games per round of Round Robin.

Therefore, the number of games in the Round Robin structure can be varied by simply increasing the number of rounds played. If k represents the number of times the format is repeated before the tournament is closed, the total number of games within the tournament is shown in equation 1.

$$k \cdot {}^{n}C_2 = \frac{kn(n-1)}{2} \tag{1}$$

where n is the number of teams in the tournament and $k \in \{1, 2, 3, \dots\}$. As a single round has a number of games proportional to n^2, increasing k to large numbers is not feasible as conducting round robin gets logistically cumbersome due to the high number of games. This fact is taken into consideration in the paper while evaluating the effectiveness of Round Robin in conveying reliable information about tournament rankings with respect to the number of games played. Teams are ranked according to win percentage.

3.2 Random Pairings of Teams

This tournament structure is exactly what the name may suggest. Teams are randomly paired and made to play with each other. It is important to note certain features about the setup of this format: the number of games played by each team is the same, and randomness is the only deciding factor of which teams pair up with each other.

Since each team plays the same number of games, it is important that the number of teams in the tournament is even. It is nontrivial to generate random pairings of teams (playing a variable number of games each) with an odd number of teams. Since the Overwatch League consists of 20 teams, this does not pose a problem.

Similar to the setup of Round Robin, the number of games each team plays is varied across multiple simulations. The number of games played in this format can be varied as $\frac{1}{2}kn$ where n is the number of teams in the tournament and $k \in \{1, 2, 3, \dots\}$. Teams are ranked according to win percentage.

3.3 The Swiss System

Of the tournament structures that are analysed in this paper, the Swiss System is the most complex. It is used widely for Chess tournaments and operates using a pairing procedure that generally pairs teams of similar abilities as shown in Hua 2017 [5]. These abilities are not known beforehand - the pairing algorithm uses wins and previous opponents to determine what teams to pair next.

This paper uses an open-source implementation of the Swiss pairing algorithm in order to conduct Monte-Carlo simulations as shown in Baker 2018 [2]. The pairing algorithm pairs teams of similar abilities while also taking care not to pair specific pairs of teams together often too much.

This is achieved by representing all teams as nodes on a graph and having weights between the graph represented by the *quality* of a pairing. Then, the **maximum weight matching** for the graph is computed: this is the set of matchings (pairings) such that the sum of the weights of the matching is maximal. This is the set of pairings with the highest quality; the quality is determined by the function that sets the weights of the pairings.

The quality of a pairing is determined by this library using algorithm 1, and the weight of a pairing is computed by algorithm 2 as detailed in the code found in Baker 2018 [2].

Algorithm 1: Function to compute the quality of a pairing

input : Importance of the pairing and Closeness between the teams being paired
output: A numerical value representing the quality of the pairing

return $(\mathsf{Importance} + 1) \times (\mathsf{Closeness} + 1)$

Algorithm 2: Weight of pairing between team i and j

input : P, a set of points accrued by each team; G, a 2D matrix of number of games played, team i, team j
output: w, weight of pairings i and j

$\mathsf{w} \leftarrow 0$;
$\mathsf{HighestGames} \leftarrow \max_k(G_{i,k})$;
if $G_{i,j} < \mathsf{HighestGames}$ **then**
  $\mathsf{w} \leftarrow \mathsf{w} + \mathtt{Quality}(\mathsf{HighestGames}, \mathsf{HighestGames}) + 1$;
end
$\mathsf{Closeness} \leftarrow \mathsf{HighestGames} - |P_i - P_j|$;
$\mathsf{Importance} \leftarrow \max(P_i, P_j)$;
$\mathsf{w} \leftarrow \mathtt{Quality}(\mathsf{Importance}, \mathsf{Closeness})$;
return w

Closeness is measured by the highest number of games minus the point differential - this is very high for teams which have very close scores. Importance is simply defined as how high the higher scoring player is - this is highest if the higher scoring player is ranked #1. This allows the Swiss pairing algorithm to match teams while prioritising good matches for higher-ranked teams. Both closeness and importance have a maximum value of HighestGames.

Another thing to note is that teams which have not played each other as many times as one of them has played somebody else will have a higher quality rating (as the algorithm adds the, highest quality possible represented by Quality(Highest, Highest) to these weights on top of computed quality) and therefore prioritised. This limits repeated matching of very closely ranked teams and ensures that there is greater diversity in the pairings.

The function that computes the quality of a pairing between two teams adds 1 to both importance and closeness to ensure that one value being 0 does not mean that the information from the other value is lost.

After the weights between each pair of teams are computed, the **maximum weight matching** of the graph is computed. There exist already researched algorithms to compute this. This algorithm will not be discussed in this paper but can be found in Osiakwan et al. 1990 [9].

After a certain number of sets of pairings, the Swiss system can be terminated, and the players can be ranked according to the number of wins they have.

This system differs from the other tournament systems as it incorporates competitive balance by design - the teams which are made to play against each other are, for the most part, very close in terms of skill. This is different from the competitive balance that is set during simulation - the Swiss system incorporates *match-level* competitive balance whereas setting the spread of the teams in the tournament refers to *championship-level* competitive balance. The result of varying championship-level competitive balance of teams going into the Swiss system on reliability is interesting to investigate and compare with the other tournament systems as there is an added layer of competitive balance on top of the match-level balancing that the Swiss system brings.

4 Algorithms and Metrics

4.1 The Elo Rating System

This paper is unconcerned with ranking teams for the Overwatch League but seeks to use data from the Overwatch League to build a probability model that underlies the generation of results from simulated tournaments. The Elo rating system provides one such probability model. It is a system which consists of teams' ratings and two simple rules that describe how those ratings are updated. It is widely used in large-scale online Chess, online video games, board games, and has been used in association with American football as well.

Matches from the Overwatch League will be processed by the Elo rating system to assign a rating to each team. The assigned Elo ratings of two teams are mathematically processed in order to calculate a probability that one team will win over another.

When two teams play each other in the Elo rating system, their ratings are used to generated an **expected performance value** for each team. The expected score of each player is given in Langville et al. 2012 [8] and is shown in equations 2 and 3.

$$E_A = \frac{1}{1 + 10^{R_B - R_A}/1000} \tag{2}$$

$$E_B = \frac{1}{1 + 10^{R_A - R_B}/1000} \tag{3}$$

R_A and R_B represent the Elo rating for teams A and B respectively, and E_A and E_B represent the expected performance value for teams A and B respectively. The constant 1000 in these formulae represents the rating of an average player in the Elo system. Although the standard value used of this constant is 400, the numerical value is arbitrary. The decision was made to set this value at 1000 in order to have greater control over the standard deviation of teams' ratings.

These expected scores are then used to update the respective scores of each team based on the outcome of the game. If S_A and S_B are values that represent the *actual performance* of teams A and B (for example, $S_A = 1$ and $S_B = 0$ if team A won over team B), equations 4 and 5 describe how the ratings of each team should be updated.

$$R'_A = R_A + K(S_A - E_A) \tag{4}$$

$$R'_B = R_B + K(S_B - E_B) \tag{5}$$

R'_A and R'_B represent the updated Elo ratings of each team. The constant K in the above equations is referred to as the K-factor. Assuming a game in which winning has an actual score of 1 and losing has an actual score of 0, this represents the maximum number of points each team can either win or lose. For the purposes of this paper, this number is set to 32, which is the standard value used in Elo calculations in most tournaments. A description of reasons why this value is widely used can be found in Langville et al. 2012 [8].

4.2 Algorithms for setting and measuring competitive balance

Competitive balance, as defined above, is a measure of the level of disparity in competitors' abilities. It is possible to quantify it through many metrics that measure inequality, although the most commonly used metric is the standard deviation of the final points of the teams as shown in Goossens 2005 [4]. Other metrics include National Measure of Seasonal Imbalance (NSMI) and the Gini coefficient, an econometric measure of income equality adapted to the sports context. However, none of these measures is used in this paper.

The most widely used measure of competitive balance is the standard deviation of teams' final scores, measuring how much spread there is in the final performance of teams. This paper uses this same idea but implements it on the ground truth ratings of the teams.

Arguably, this is more effective than calculating the standard deviation of the final scores of teams at the end of each tournament, as it allows the separation of two different notions: how much spread the tournament system itself creates due to processes that occur within it, and how spread apart the teams are in their *inherent abilities* against each other.

For setting competitive balance, a randomised method is used. Since competitive balance is varied in terms of the standard deviation of the ratings of the teams, the rating of each team was set using a random gaussian value with mean 1000 and standard deviation in the following range:

$$\sigma \in \{100, 200, 300, 400, 500, 600, 700, 800, 900\}$$

There were nine different sets of 20 teams each of which had their own standard deviation set using a value in the above range. Although the values sampled from the Gaussian function might not have a standard deviation very close to the σ used to generate them, the sampling process is repeated until this value is within $\sigma \pm 20$.

4.3 Simulating the tournament

When a set of teams is generated for a specific level of competitive balance, matches are allocated according to each of the three tournament systems being investigated. A probabilistic method is used to decide the outcome of a match between two teams.

The formulae for expected performance values are given in Section 4.1. For the purposes of this paper, there are no scores incorporated into the matches. Since winning has an actual score of 1 and losing has an actual score of 0, the expected performance values (which is a prediction of the actual score of a team) provides the win probabilities of each team. Note that for two teams A and B, $E_A = 1 - E_B$.

These win probabilities are then used to pick a winner. For example, if the win probability for team A is 0.57, a continuous random variable is generated and compared to 0.57. If it is less than or equal to 0.57, team A is declared as the winner. Otherwise, team B is declared as the winner.

4.4 Comparing rankings

In this paper, three methods of comparing rankings are used.

1. Spearman's Rank Correlation Coefficient
2. Kendall's Tau
3. Levenshtein Distance

4.4.1 Spearman's Rank Correlation Coefficient and Kendall's Tau

The first two metrics used to compare rankings are very similar and are widely used in the field. Both of them are used to measure the strength and direction of the association between two ranked variables. They equate to 1 if the relationship between the two variables is **monotonic**, that is, as the value of one variable increases, so does the value of the other variable.

Spearman's correlation has its similarities with mean squared error, as notable from its definition given by following formula in Kendall 1970 [6] in equation 6.

$$r_s = 1 - \frac{6 \sum (A_i - B_i)^2}{n(n^2 - 1)} \tag{6}$$

where A and B are sequences in which the rank of team i is represented by A_i. This formula only computes the Spearman coefficient for integer rankings which do not have ties within them. As this is the case within the paper, this formula is used.

This outputs a value between 0 and 1 representing the correlation between two sets of ranks A and B.

Kendall Tau distance, in words, computes the pairwise disagreements between two ranks A and B. A disagreement is a situation in which both two elements are ordered oppositely in each ranking. Mathematically, it is defined in Kendall 1970 [6] and in Fagin et al. 2003 [3] in equation 7.

$$K(A,B) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(A,B) \tag{7}$$

where:

1. P is the set of unordered pairs of distinct elements in A and B
2. $\bar{K}_{i,j}(A,B) = 0$ if i and j are in the same order in A and B
3. $\bar{K}_{i,j}(A,B) = 1$ if i and j are in the opposite order in A and B

Since Kendall Tau distance is a metric that computes pairwise disagreements, and there can be made exactly $nC_2 = \frac{n(n-1)}{2}$ pairs in any two rankings, it is normalised by dividing by $\frac{n(n-1)}{2}$.

Since a value of $2\frac{K(A,B)}{n(n-1)}$ indicates maximum disagreement between two sets of rankings, subtracting that from 1 as shown above reports the *correlation* between two sets of data - a higher value indicates higher correlation. Therefore, when this paper reports Kendall-tau, the value that will be reported is shown in equation 8.

$$K(A,B)_{\text{normalised}} = 1 - 2 \cdot \frac{K(A,B)}{n(n-1)} \tag{8}$$

4.4.2 Levenshtein Distance

Levenshtein distance is a measure of **edit distance** between two sequences. There are many existing measures and interpretations of edit distance, one example being Hamming distance. However, Levenshtein is defined as the minimum number of edits (i.e. insertions, deletions, or substitutions) required to change one sequence into another. Although it is used widely to quantify the "closeness" of two strings, earning the classification into the group of existing string similarity algorithms, it is used here in order to quantify the closeness of two rankings of teams.

It can be computed using a dynamic programming algorithm based on the recurrence in Yujian et al. 2007 [11] and Babar 2020 [1] which is reproduced in equation 9.

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j), & \text{if } \min(i,j) = 0. \\ \min = \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & \text{otherwise.} \end{cases} \tag{9}$$

$1_{a_i \neq b_i}$ is 0 when $a_i = b_i$ and is 1 otherwise. $\text{lev}_{a,b}(i,j)$ represents the Levenshtein distance between the first i characters of sequence a and the first j characters of sequence b. So, the Levenshtein distance between the whole strings a and b is given by $\text{lev}_{a,b}(n,n)$ where n is the length of each string. For the purposes of comparing rankings, it is assumed that the length of the strings (the number of teams in each ranking) is the same.

An intuitive explanation in words of the recurrence used in equation 9 is as follows: when either of i or j are zero, the minimum edit distance between the two substrings of a and b is simply the length of the longest substring. This is the base case presented by the recurrence in equation 9. Otherwise, it is the minimum of the three following values:

1. One more than the edit distance between the first $i-1$ characters of a and the first j characters of b
2. One more than the edit distance between the first i characters of a and the first $j-1$ characters of b
3. The same as the edit distance between the first $i-1$ characters of a and the first $j-1$ characters of b **only if the ith character of a and the jth character of b are the same**. Otherwise, it is one more than $\text{lev}_{a,b}(i-1,j-1)$.

Choosing the minimum of the three values above results in the minimum number of insertions, substitutions or deletions required to transform one string into another. Thus, this recurrence can be used to compute Levenshtein distance.

The maximum value of the Levenshtein difference between two strings would be the maximum of the length of the two strings. Thus, it is possible to normalise Levenshtein distance between 0 and 1 as is done in Kendall-tau by first dividing the value of $\text{lev}_{a,b}(n,n)$ by n and then subtracting it from 1. This is done in equation 10.

$$\text{lev}_{a,b}(n,n)_{\text{normalised}} = 1 - \frac{\text{lev}_{a,b}(n,n)}{n} \tag{10}$$

The first two metrics were chosen to report the correlation between two rankings while the third was chosen to report closeness. All three metrics are normalised, and so will appear as values between 0 and 1.

Each of these metrics are run multiple times, and the final number that represents closeness for a specific set of tournament rankings over T trials is given in equation 11.

$$\text{reliability} = \frac{\Sigma_{k=1}^{T} f(r_{\text{true}}, r_k)}{T} \tag{11}$$

where $f(a,b)$ is the function used to compare rankings.

5 Results

5.1 Competitive balance and reliability

Results were obtained for random pairings, round-robin tournaments, and for Swiss pairings. The reliability of the results are measured as shown above and are plotted against the number of games played in each tournament system to investigate how reliability changes based on how much information is fed into the tournament system in the form of a game between two teams. The first set of teams (which have $\sigma = 89$) are real teams from the Overwatch League; the rest are generated data. The legend for curves representing sets of teams with different competitive imbalances is given in figure 1.

It is clear that the reliability of the rankings produced from each tournament generally increases as the competitive imbalance, represented by the standard deviation in the teams' ratings, rating σ, increases. This is observable across all three tournaments in figures 2, 3 and 4 and all three metrics shown in those figures. However, it is never the case that the set of teams with the most competitive imbalance (highest σ) produces the most reliable rankings - there is always a point after which the marginal benefit of increasing the competitive balance is close to 0. In other words, the metric/number of games curves for higher values of σ converge across all tournament structures and metrics.

This could be due to bottlenecks that exist within the tournament structure itself that need to be surpassed before increasing competitive imbalance has the same effect it induced initially. These bottlenecks could be the number of teams, whether the scores of the winners and losers are incorporated, the way the tournament is set up, or the inherent randomness in the results of a match. The uncertainty due to the inherent randomness may be overcome by increasing the number of games, but this is infeasible past a certain threshold.

Figure 1: Legend for curves representing sets of teams with different competitive imbalances (σ)

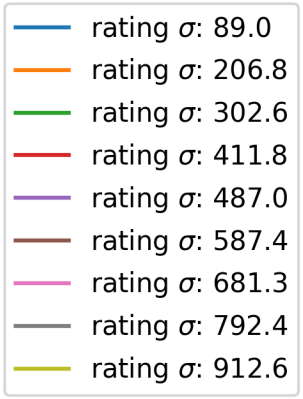
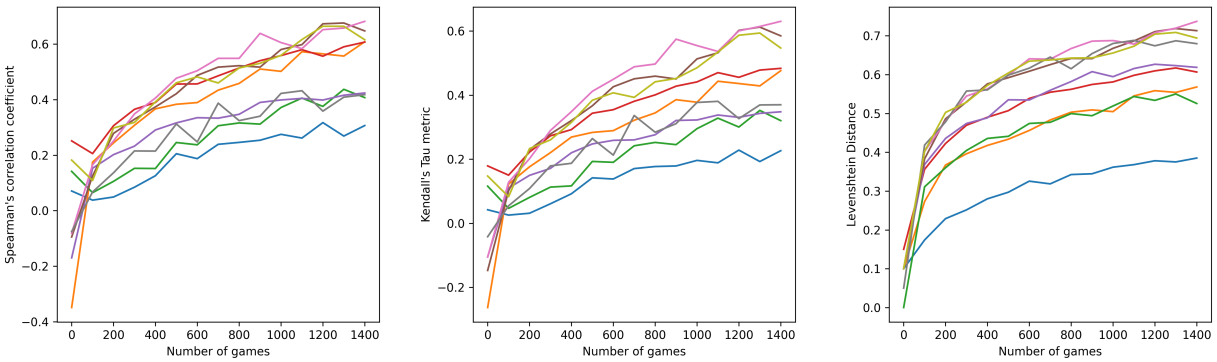


Figure 2: Spearman-r, Kendall Tau, and Levenshtein Distance for rankings in random pairings of teams versus number of games played



5.2 Comparing tournament structures

The figures shown in section 5.1 are represented differently in order to facilitate comparison between tournament structures. The reliability/games curves for each tournament that is held for the same set of teams are placed within the same figures for each metric. The legend for these figures is present in figure 5.

At lower levels of competitive imbalance ($\sigma = 89, 206$) as shown in figures 6 and 7, there is a relatively high level of fluctuation in all the metrics that represent reliability. However, as competitive imbalance increases in the range $\sigma = 302, 411, 487$ as shown in figures 8, 9 and 10 respectively, the reliability curves of each tournament are more consistent. This relationship is observed even further in figures 11, 12, 13, and 14: the curves smoothen out and remain relatively consistent. This is because the level of tournament outcome uncertainty falls as the level of competitive imbalance increases. Further work needs to be conducted to investigate tournament outcome uncertainty. More details are present in Section 6.

There is a general observation across all figures: round-robin tournaments are the most reliable at any given point, random pairings are the second most reliable, and Swiss pairings are the least reliable out of the three. This result is consistent with Hua 2017 [5] in which the Swiss system performed more poorly than a simple random pairing of teams. The difference in reliability increases as the competitive imbalance increases - as the teams become more spread out in terms of skill level, round-robin is seen to be much more reliable than both random pairings and Swiss pairings, and random pairings are seen to be more reliable than Swiss pairings. This effect is shown by the fact that there is little separation between the curves in figures 6 and 7, after which the curves begin to diverge dramatically in figures 8 and 9. This divergence continues to increase for sets of teams with higher competitive imbalance as shown in figures 10, 11, 12, and 13.

Figure 3: Spearman-r, Kendall Tau, and Levenshtein Distance for rankings in swiss pairings of teams versus number of games played

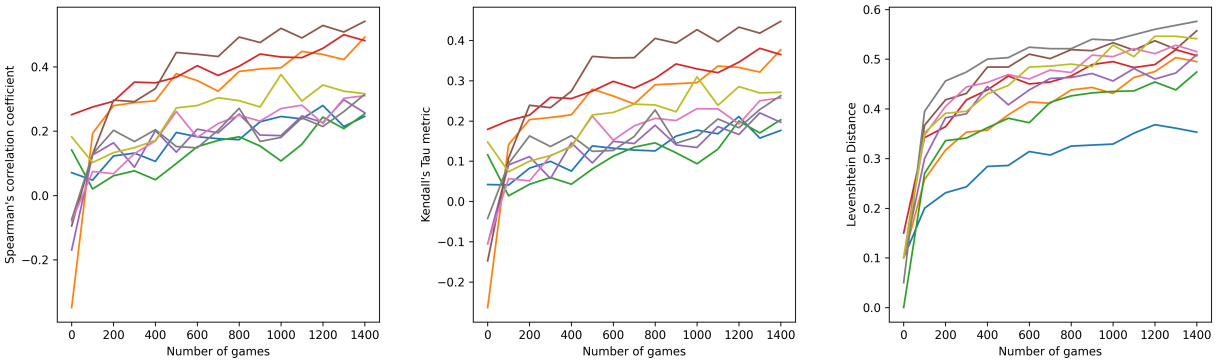
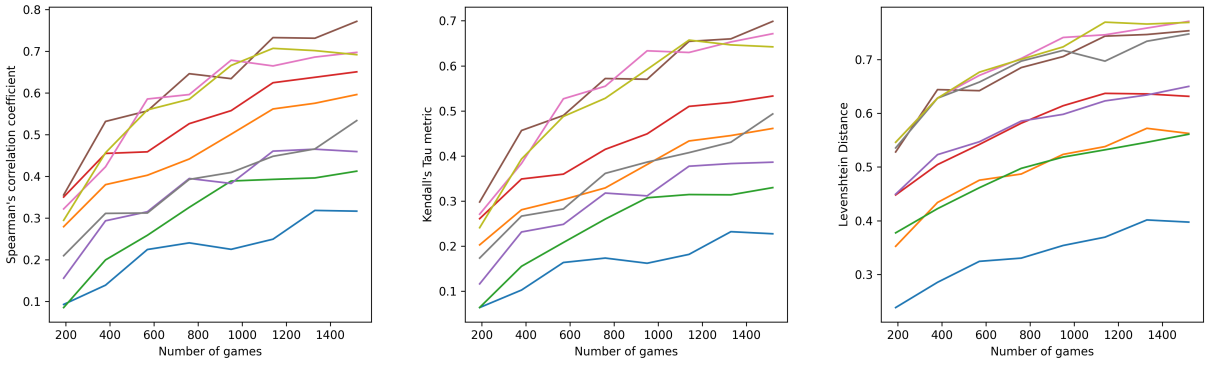


Figure 4: Spearman-r, Kendall Tau, and Levenshtein Distance for rankings in round robin tournaments of teams versus number of games played



This does not indicate that the round-robin format is superior in all regards to the other tournament structures - although it consistently outputs rankings closer to the ground truth, this may not be the priority of a tournament administrator. Reliability is only one of many things to be desired from a successful tournament. Other factors include the level of viewership and interest (which directly translates into revenue for many sports and eSports tournaments) and the possibility for any modest team to win.

Although the Swiss system is less reliable than the other structures, a possible reason for its popularity is the fact that it allows for interesting games between teams. Every team meets opponents of roughly the same ability and skill level; the win percentage of a team might not be as indicative of its ability as it is in a round-robin tournament for this very reason. Thus, using win percentage to measure teams' abilities is less meaningful than in other tournament systems.

Both of these factors are linked to competitive *balance*. A more competitively balanced set of teams increases the probability for a less able team to win the championship, which can sometimes be a priority for tournament administrators as it provides an incentive for less competitive teams to enlist in a tournament and for viewers who are affiliated with those teams to follow the tournament for longer. Although it is very clear that round-robin can be relied on the most to produce rankings most representative of teams' abilities, further work in tournament outcome uncertainty may illuminate a decision making strategy for tournament administrators with different priorities in choosing a tournament structure to meet their needs.

Figure 5: Legend for curves representing tournament structures

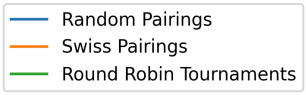


Figure 6: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 89$

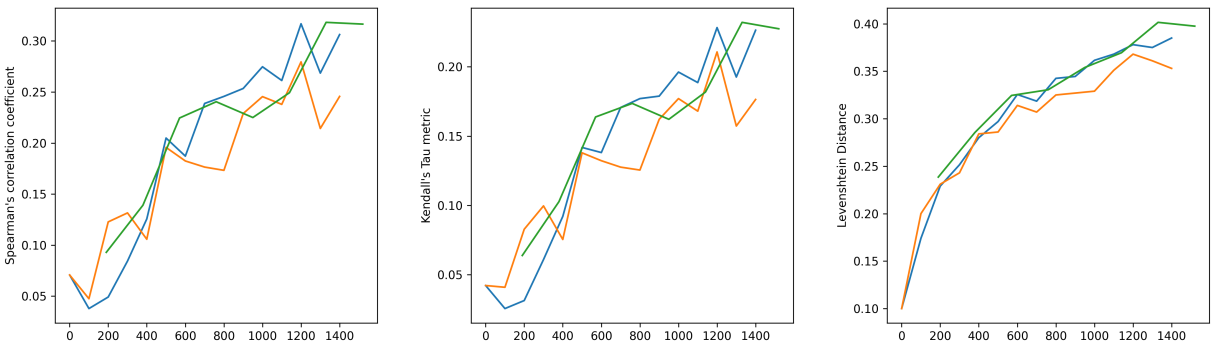


Figure 7: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 206.8$

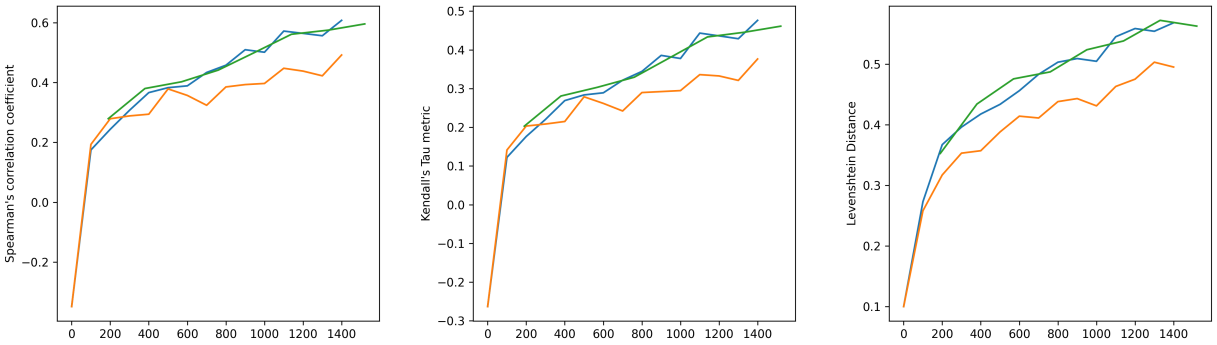


Figure 8: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 302.6$

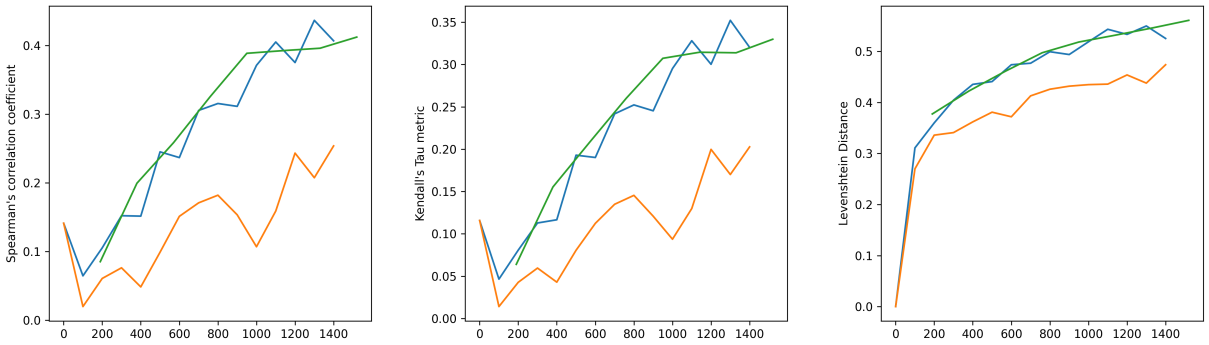


Figure 9: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 411.8$

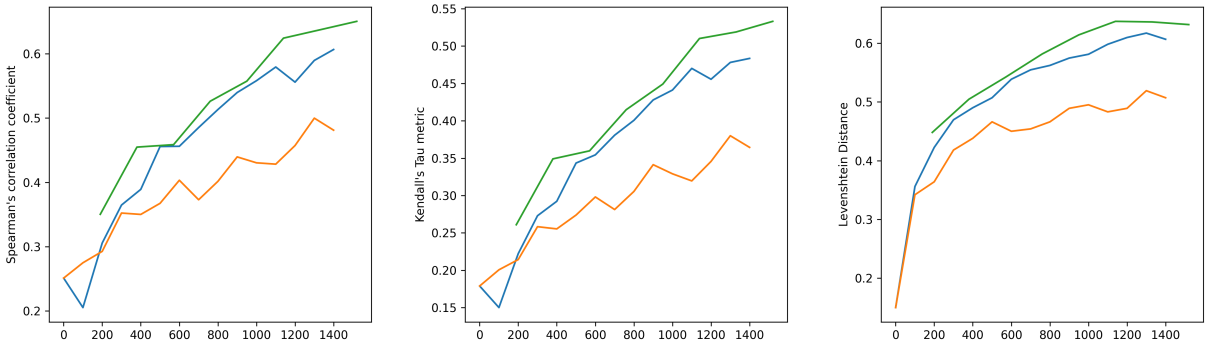


Figure 10: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 487.0$

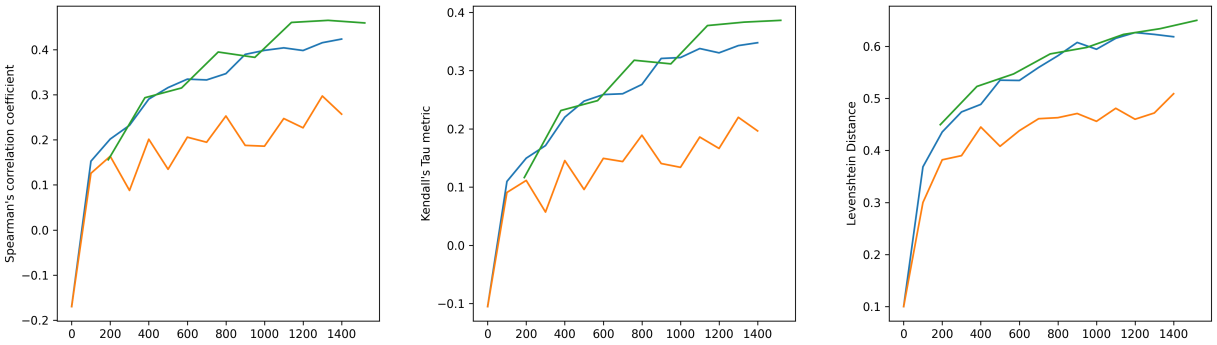


Figure 11: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 587.4$

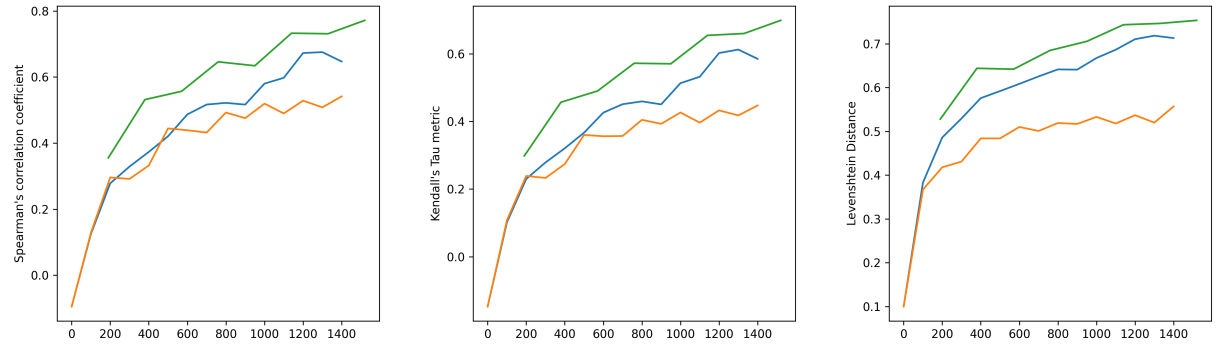


Figure 12: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 681.3$

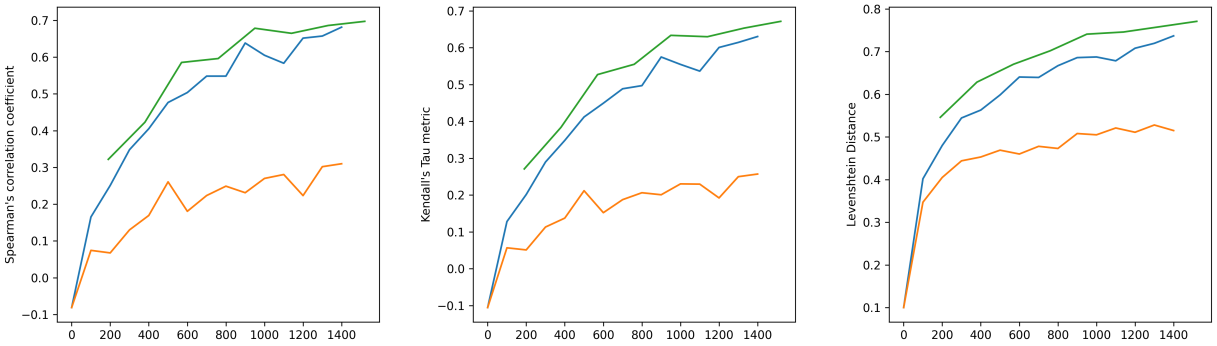


Figure 13: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 792.4$

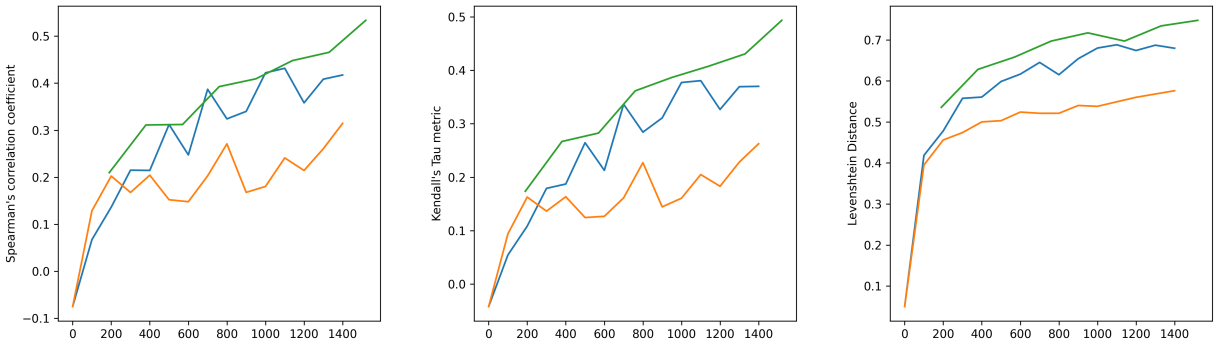
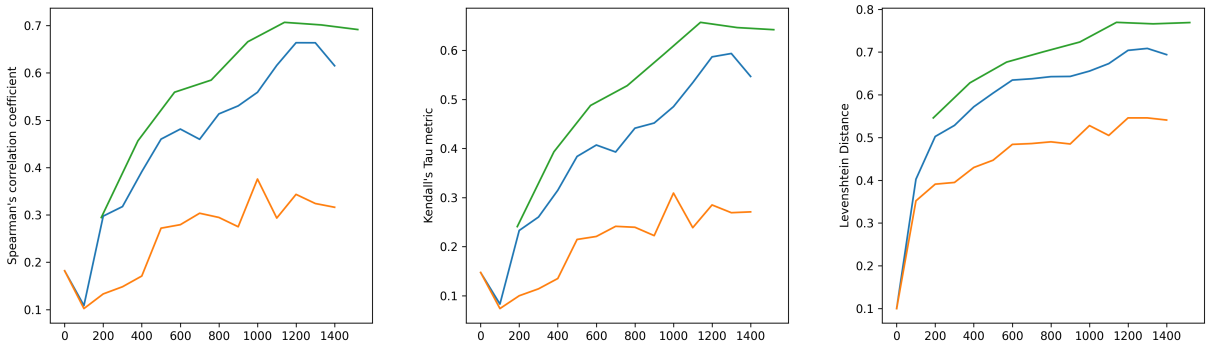


Figure 14: Spearman-r, Kendall Tau, and Levenshtein Distance for all tournaments with $\sigma = 912.6$



6 Further Work

The relationship between competitive imbalance and reliability is most strongly observed in the Levenshtein distance metric. A reason for this could be that this metric, unlike Kendall-tau, measures *closeness* rather than correlation. Although all of these metrics can be formulated in terms of distance, Kendall-tau measures the number of *disagreements* in the two rankings whereas Levenshtein measures the least number of insertions, swaps or deletions required to transform one ranking into another without any regard for the actual ranks which differ. Both Kendall-tau and Spearman take this into account. Further work can be done to investigate the differences between the three metrics, and more importantly, the merits and demerits of using a metric that represents minimum edit distance like Levenshtein rather than correlation between two rankings. These merits can be investigated in the context of reporting reliability as is done in this paper.

Even in structures which have fluctuating values for Spearman's coefficient and Kendall-tau as the number of games increases, the curves for Levenshtein distance are much smoother and have more consistent relationships. Another parameter that could be incorporated into Levenshtein distance is prioritization of higher ranked teams: all the metrics used in this paper did not account for the fact that a ranking which is correct for higher ranked teams can be considered more reliable than one which is only correct for lower ranked teams. Thus, a weighted Levenshtein distance metric is also proposed.

It was noticed in this paper that the Swiss system performed worse than both round-robin and random pairings in terms of the reliability metric defined. It might be possible to adapt the method of ranking within the Swiss system (which is currently ranking by win percentage) to weight the win percentages by strength of schedule after the tournament has finished, producing a measurement that more accurately reflects their true abilities. As the Swiss system was designed to result in more "interesting" games scheduled between teams of similar abilities, another alternative metric can be explored that measures this. Although this may seem similar to a simple measure of competitive balance within the tournament system, an account of the strength of schedule in the matches conducted in the tournament may prove to be helpful in computing this metric.

One major aspect of a tournament that is related to both reliability and competitive balance is tournament outcome uncertainty. Although studies have been done on the effect of competitive balance on tournament outcome uncertainty, there is value in unravelling the relationship between reliability and tournament outcome uncertainty in the context of competitive balance. The fact that reliability curves are smoother at higher levels of competitive imbalance indicates a positive relationship between competitive balance and tournament outcome uncertainty. More work can be done to investigate the specific nature of this relationship.

Another area of work that is proposed is into the specific factors that bottleneck reliability at a certain point, conferring no additional benefit when competitive imbalance is increased. This is possible by an investigation into both the parameters used to set up tournaments while also incorporating a study of tournament outcome uncertainty. One major assumption of this paper is that the abilities of teams follow a normal distribution - this might not be the case in certain scenarios, and further work can be done to determine whether the reliability curves change if teams' abilities are not normally distributed.

Acknowledgement The author thanks Dr Tim Chartier of Davidson College for advising this work and the Pioneer Research Program for the opportunity to pursue research under his guidance.

References

[1] Nikhil Babar. The levenshtein distance algorithm - dzone big data. `https://dzone.com/articles/the-levenshtein-algorithm-1#:~:text=TheLevenshteindistanceisa,onewordintotheother.`, Aug 2020.

[2] Baker Thomas David. Swiss pairing algorithm. `https://github.com/bakert/swiss`, Apr 2018.

[3] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[4] Kelly Goossens. *Competitive balance in European football: Comparison by adapting measures: National measure of seasonal imbalance and top 3*. University of Antwerp, Research Administration, 2005.

[5] Christopher Hua. The Swiss Tournament Model. 2017.

[6] M.G. Kendall. *Rank Correlation Methods*. Theory and applications of rank order-statistics. Griffin, 1970.

[7] M. Kringstad and Bill Gerrard. The Concepts of Competitive Balance and Uncertainty of Outcome. IASE Conference Papers 0412, International Association of Sports Economists, May 2004.

[8] Amy N Langville and Carl D Meyer. *Who's# 1?: the science of rating and ranking*. Princeton University Press, 2012.

[9] Constantine NK Osiakwan and Selim G Akl. The maximum weight perfect matching problem for complete weighted graphs is in pc. In *Proceedings of the Second IEEE Symposium on Parallel and Distributed Processing 1990*, pages 880–887. IEEE, 1990.

[10] Philip Scarf, Muhammad Mat Yusof, and Mark Bilbao. A numerical study of designs for sporting contests. *European Journal of Operational Research*, 198(1):190–198, 2009.

[11] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.