DECISION SCIENCES INSTITUTE

Assessing the Convergence of the Elo Ranking Model

Thomas R. Robbins
East Carolina University
robbinst@ecu.edu

ABSTRACT

The Elo rating system is a probabilistic skill ranking algorithm originally developed for chess. It has since gained widespread adoption and been applied to games such as Go, backgammon, Scrabble, and various video games. Elo-based models have also been adapted to physical sports like tennis, football, basketball, and baseball, as well as numerous eSports. In this paper, we review the Elo system and its applications. We develop a simulation model that examines how the model's rantings converge toward true skills. With enough matches the model can converge to an accurate ordinal ranking, but estimated win probabilities remain error prone.

KEYWORDS: Elo Rating System, Ranking Algorithms, Power Ranking Models, Simulation

INTRODUCTION

The Elo rating system is a method for estimating the relative skill level of competitors in zero sum competition (Wikipedia 2025; Elo 1978). The system was developed by physics professor Arpad Elo as a method for rating chess players and was adopted by FIDE (the International Chess Federation) in 1970 as the standard ranking system for ranking chess players internationally. Elo published the details of the system to a wider audience in book form in 1978.

The Elo system has become quite popular not only as a chess rating system, but as a system to rate competitors in other games such as Go, Scrabble, and Backgammon. Elo models have also been adopted to video games and eSports, as well as competitive sports such as tennis, soccer, American football, and basketball.

In an Elo system each competitor has a rating, expressed as a point value. The estimated probability a competitor defeats another competitor is based on the point differential. The Elo system is a zero-sum system; the winner of a competition takes points from the loser. The magnitude of the point transfer is based on the relative skill levels of the competitors and a scaling constant *K*. The point transfer is based on the probability of victory. A competitor takes few points from a competitor they were heavily favored to win, but they take many points from a winning a competition they were projected to lose. The scaling constant determines the maximum number of points that can be transferred from a single competition.

The Elo ranking model serves multiple purposes. It gives each competitor an absolute score, it rank-orders competitors based on those scores, and it calculates the probability of victory for any two competitors based on their relative scores.

While widely used the Elo model is often criticized. The model assumes stationary skill levels yet individuals in games such as chess become better with experience, or worse as their skills deteriorate. In team sports injuries and change in personnel can dramatically shift the

underlying skill level of the team. A competitor's rating can be highly dependent on the ordering of the competition. The basic Elo model only looks at the final outcome; it ignores the margin of victory and makes no allowances for home field advantage. In team sports the standard Elo model makes no allowance for the significant changes that can occur in between seasons.

Most of the research related to Elo models addresses modifications to the model that may make it more effective. None of literature addresses the fundamental question we ask – if the assumptions of the model are correct how accurate is the model under various conditions, and for various applications. Does the model provide an accurate skill level? How accurate is the ranking order? Is the win probability accurate enough to inform betting odds calculations?

In this paper we assume that the basic assumptions of the Elo model are correct. We assume that every competitor has a true skill level that is estimated by the Elo score. Furthermore, we assume that the difference in skill levels is an accurate probabilistic forecast of victory between any two competitors. Under these ideal conditions we seek to evaluate the accuracy of the Elo predictions over time. We examine the accuracy in the overall skill rating of each competitor, the accuracy of relative ranking of the competitive field, and the accuracy of the estimate win probability in each match up.

LITERATURE REVIEW

The Elo model was first publicized in Elo (1978). Elo presents the model, detailing the statistical concepts and reasoning of the model in the context of chess. He uses the model to analyze the performance of various chess masters. Since then, the model has been discussed in literature with authors offering critiques, applications, and extensions.

One stream of literature focuses on the application of the Elo model to new environments. Carbone, Corke, and Moisiadis (2016) use an Elo based model to predict the outcome of rugby matches. Tenkanen (2019) applies Elo to NHL games, while Chen, Kok, and Heiser (2018) applies Elo to women's soccer. Bigsby and Ohlmann (2017) evaluate against other models and find it to be a good predictor for collegiate wrestling. Wolf et al. (2021) develop an Elo like model to rate individual soccer players. Lehmann and Wohlrabe (2017) adopt the Elo model to competition among economics journals. Journals compete each year for impact resulting in a win or loss and an Elo based journal ranking.

Vaziri et al. (2018) compare Elo to four other ranking models against three fairness and comprehensiveness criteria: accounting for opponent strength, always rewarding wins, and being independent of match order. They criticize Elo for its dependency on match order, but find that none of the models evaluated meet all three criteria. Kovalchik (2016) examine multiple models for predicting tennis matches and finds that Elo models perform well. Vaughan Williams et al. (2019) also evaluate multiple models in tennis, and also find good performance from Elo based models. Chen, Kok, and Heiser (2018) develop an Elo based system for women's UEFA soccer. Robbins (2023) examines the accuracy of the Elo model's probabilistic prediction for NFL Games relative to other power index models and betting odds.

Much of the literature is focused on enhancements to the standard Elo model. Hvattum and Arntzen (2010) use Elo rating differences as the covariates in an order logit regression model. They find the model performs well given its simplicity, but is outperformed by betting odds. Szczecinski (2020) develops a generalized version of the Elo model that accounts for margin of victory. Kovalchik (2020) introduces a similar margin of victory adjustment. Angelini, Candila, and Angelis (2021) propose a weighted Elo model for tennis that also factors in the margin of

recent victories. Langholf (2018) modifies the Elo score adjustment to create a self-justifying Elo score. Morrison (2019) compares Elo with other algorithms for gaming applications. Aldous (2017) analyze the underlying statistical properties of the Elo model and use simulation to evaluate how well it responds to changes in team strength.

APPLICATIONS OF THE ELO MODEL

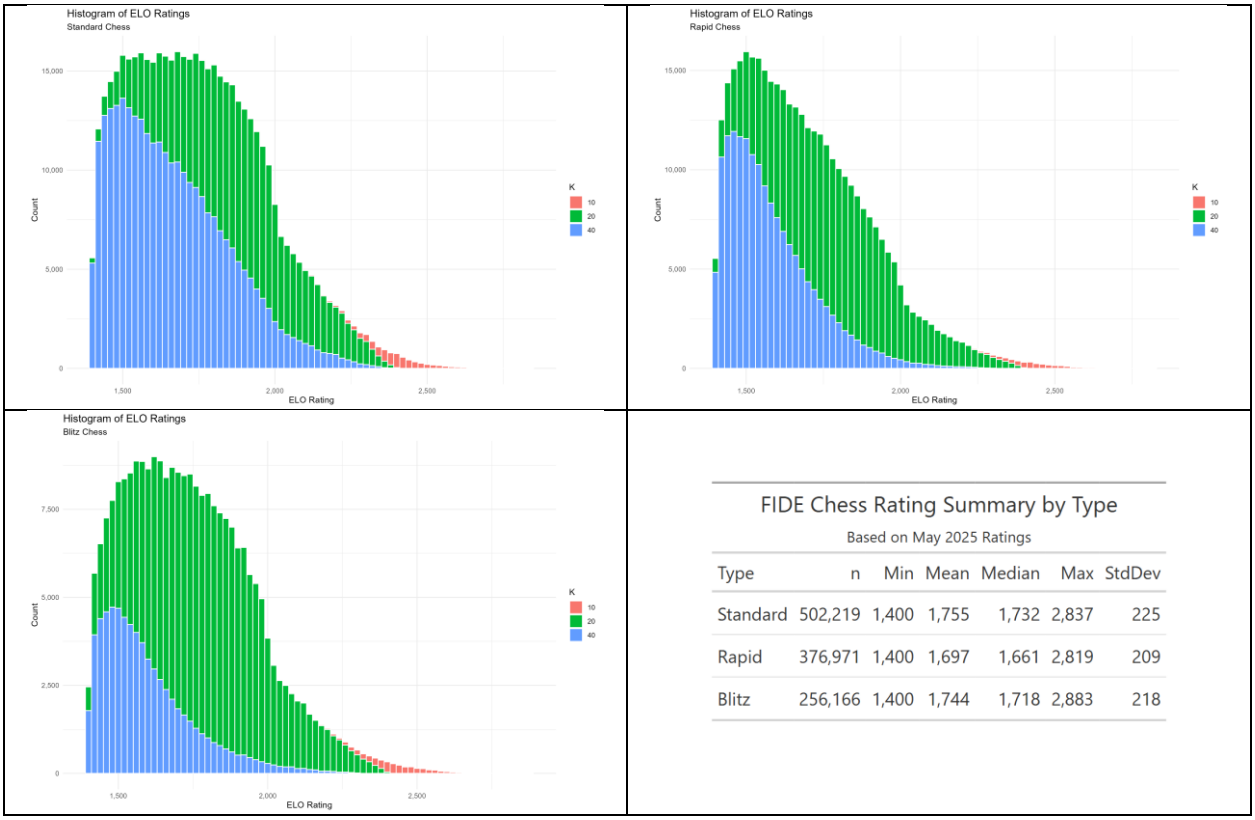
The Elo model has been utilized in a wide range of competitive environments. In this section we review the breadth of environments where versions of the Elo model have been applied and examine the properties of the associated data sets.

Games

Chess is the original application of the Elo model, and it remains an important aspect of competitive chess. The International Chess Federation (FIDE) maintains and publishes a list of Elo based rankings for registered FIDE players (FIDE 2025). Only competition in FIDE rated tournaments counts toward the rating, and only players with a rating of at least 1400 are published on the site. FIDE publishes ratings for Standard, Rapid, and Blitz chess.

Figure 1 provides a summary of the Elo scores across the three forms of chess that are rated. Several things are worth noting. First because only players with a minimum score of 1400 are listed, the distribution of rankings is left censored. The remaining distribution appears highly skewed, though if low rated ratings were included it would be much less so. Initial rankings in the FIDE system are assigned only after players play at least five games against other players, and then the initial ranking is developed as an average of the opponents' rankings adjusted based on the win rate. Initially players' rankings are updated based on the adjustment factor K of 40, to allow for rapid adjustment of the rating. Once a player has played 30 rated games the K factor will be reduced to 20. If a player reaches a rating level of 2400, the K factor is permanently reduced to 10, even if the rating subsequently drops below 2400.

Figure 1- FIDE Elo Scores



FIDE Chess Rating Summary by Type

Based on May 2025 Ratings

Type	n	Min	Mean	Median	Max	StdDev
Standard	502,219	1,400	1,755	1,732	2,837	225
Rapid	376,971	1,400	1,697	1,661	2,819	209
Blitz	256,166	1,400	1,744	1,718	2,883	218

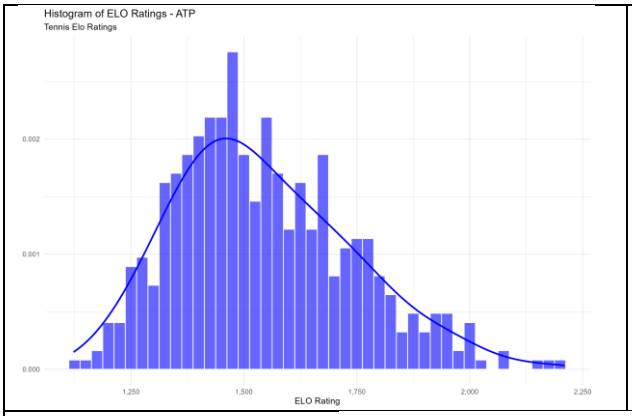
Elo based rating systems are used in other games including Go (Goratings.org 2025), Scrabble (ScrabbleAustralia 2025) and the video game Warhammer (StatCheck 2025). Go Ratings presents ratings for 926 Go players with ratings from 2390 to 3859. The Scrabble Australia site has ratings for 476 players. The ratings are based on a median of 552 games, with a maximum of 8,474 games. The Warhammer site includes rankings on nearly 30,000 players with Elo scores ranging from 1185 to 2068. Players are added to this ranking with an initial Elo of 1500.

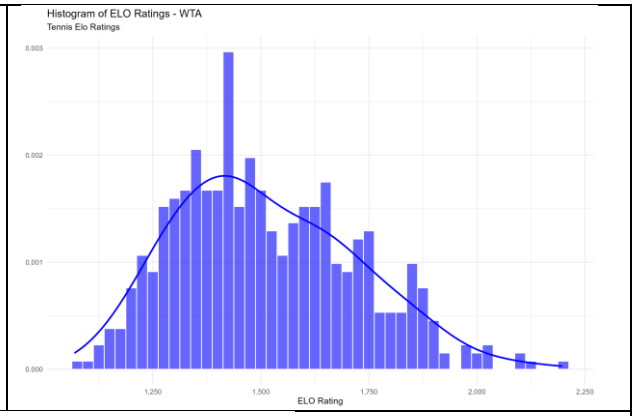
Individual Sports

Elo based rankings have been used in a number of individual games and sports, including billiards (Fargogate 2025), badminton (Badmintoncentral 2010) and table tennis (RacketInsight 2025). One sport where Elo-based systems have been widely applied is Tennis. The website tennisabstract.com maintains an Elo based rating system for both men, Association of Tennis Professionals (ATP), and women, Women's Professional Tennis (WTA) (tennisabstract.com 2025b, 2025a).

Figure 2 shows the Elo ratings for ATA and WTA players from Tennis abstract. This site only lists a current rating and with a relatively low number of players the histogram appears irregular, so a kernel density has been added. Each ranking set has a mean near 1500, but both are positively skewed with top rankings near 2200.

Figure 2 - Tennis Elo Scores





Tennis Rating Summary by Type

Based on May 2025 Ratings

Type	Sex	n	Min	Mean	Median	Max	StdDev
WTA	F	526	1,068	1,511	1,480	2,197	206
ATP	M	493	1,123	1,544	1,514	2,210	194

Team Sports

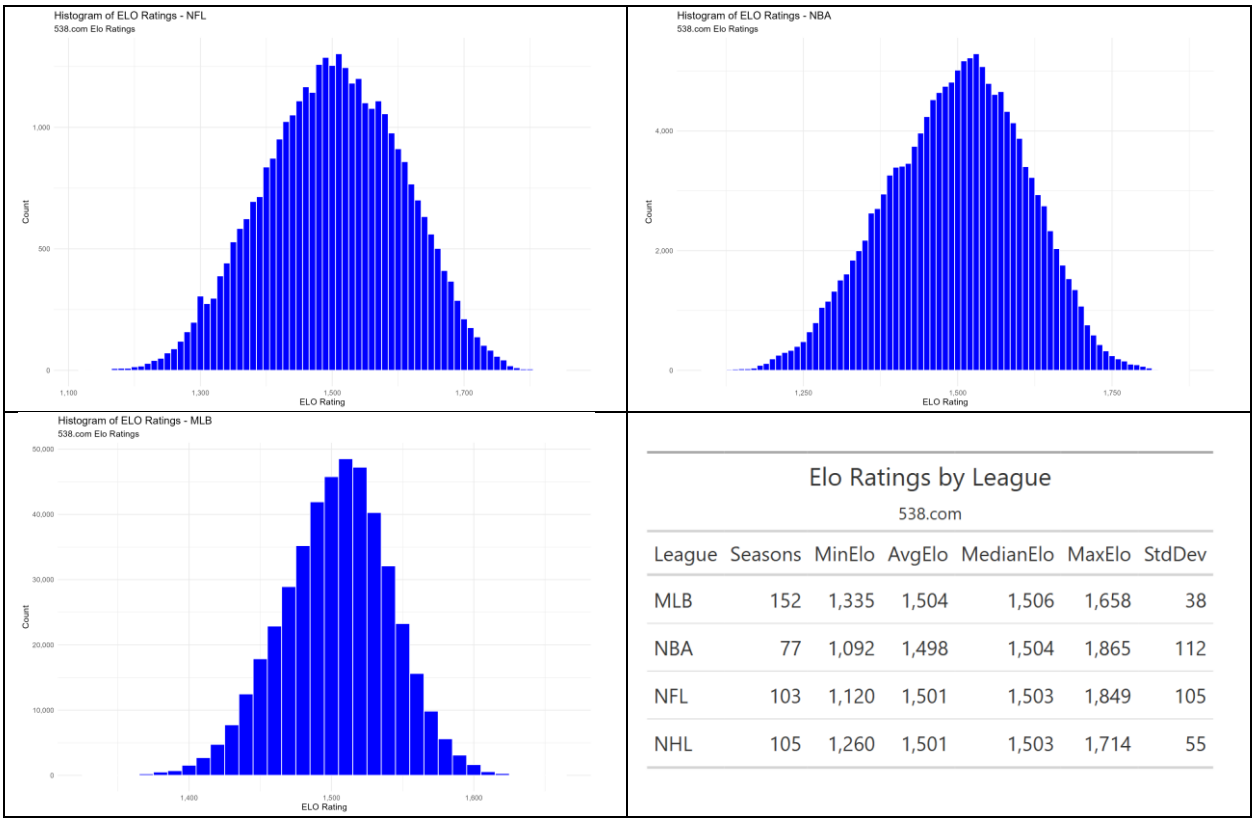
Elo based systems have been applied in a wide variety of team sports, including by the now defunct FiveThirtyEight .com sports analytics group. FiveThirtyEight.com was highly transparent about how there ranking system worked, providing detailed explanations that are still accessible (Nate Silver 2023; Silver 2014). FiveThirtyEight made several extensions to the standard Elo model. In football, when computing win probabilities they adjusted the Elo rankings to adjust for home field advantage to account for away team travel distance. They include a win differential factor to the point transfer, and they present a second model that accounts for the team's quarterback. The most significant enhancement is how they deal with the season effect in team sports. The skill level of a team can change significantly during the offseason due to change in team personnel, for either better or worse. The team that starts a season may be very different from the team that ended the last season so carrying over the Elo rating is problematic. FiveThirtyEight makes two major adjustment, one of which is very transparent, the other is much less so. First, they regress the rating toward a mean of 1505 with a factor of one-third. So for example if a team ended the prior season with a rating of 1535, the offseason regression would be 10 points, one-third the difference between the team's rating and the 1505 mean. The far less transparent adjustment relies on published over-under betting odds, "converting over-under expected wins to an Elo scale." FiveThirtyEight comments that this adjustment "helped significantly improve predictive accuracy in backtesting" (Nate Silver 2023).

While the FiveThirtyEight.com Elo ratings are no longer updated, historical data is available online (Kaggle 2021c, 2021b, 2021a). Elo rankings are also available for college football from the cfbfastR library (cfbfastR 2025).

In Figure 3 we show details on the Elo rankings for the major North American team sports. The data sets include weekly ratings for each team, generally starting from the initiation of the league so the data sets are quite large. Each graph is roughly bell shaped and has a mean very close to the 1505 value. The range of Elo scores is notably narrower in the NHL and MLB—

leagues characterized by higher game-to-game variance and greater influence of chance on outcomes, and much higher in the NBA. The range of NFL scores is in between. FiveThirtyEight.com uses a *K* value of 20 for the NFL, and slightly smaller values for other sports.

Figure 3 - Team Sports Elo Score from 538.com



Elo Ratings by League

538.com

League	Seasons	MinElo	AvgElo	MedianElo	MaxElo	StdDev
MLB	152	1,335	1,504	1,506	1,658	38
NBA	77	1,092	1,498	1,504	1,865	112
NFL	103	1,120	1,501	1,503	1,849	105
NHL	105	1,260	1,501	1,503	1,714	55

ELO MODEL DETAILS

In the Elo ranking system every competitor is given a numerical score. Competitions are evaluated by comparing the scores of the two competitors, resulting in a probabilistic forecast. The forecast for each competition is expressed as a win probability. The base probability that competitor *A*, with Elo score R_A will defeat competitor *B*, with Elo score R_B is given as

$$\Pr(A) = \frac{1}{10^{\frac{R_B - R_A}{400}} + 1} \tag{1.1}$$

After the competition Elo points are adjusted based on the outcome, in a zero-sum exchange where the winner takes points from the loser. The base level of points transferred is derived based on how likely the competitor was to win based on the pregame Elo ratings. If E_A is the probability competitor *A* would win, S_A is 1 if competitor *A* won, 0 otherwise and *K* is a scaling constant, then the new Elo rating for competitor *A* is

$$R_A^{'} = R_A + K(S_A - E_A) \tag{1.2}$$

Each competitor's Elo level will evolve over time, changing with the results of each competition. A competitor's Elo score is an important and often cited metric. Chess player Magnis Carlson is

well known for having the highest Elo rating ever for a chess player, 2882, some 31 points above that of Gary Kasparov (Wikipedia 2025). But a more fundamental consideration is the relative ratings of each player. Relative ratings will define the rank ordering of competitors and will determine the forecasted win probability when two competitors face off. So, from an evaluation perspective we should expect Elo ratings to be close to the true skill level for all competitors, for Elo based rankings to be consistent with the true skill based ranking, and for the win probability assigned to a competitor to be close to the true win probability.

THE ELO SIMULATION MODEL

To evaluate the Elo model, we developed a simulation model. At the most fundamental level, we assume the Elo model is correct. By that we mean that every competitor has a true skill level, and their performance is normally distributed around that level. The probability that competitor A defeats competitor B is given by equation (1.1). The true skill level is the actual skill level of the competitor used to calculate win probabilities according to the Elo model. The Elo rating is the estimated skill level currently assigned by the ranking model. The Elo rating is the estimate of the skill level.

Our model is an agent-based simulation. Each competitor is assigned a skill level at initialization time. We assume that skill level is fixed and unchanging. Our model allows for multiple methods for assigning the initial Elo rating. It can be set to a constant value, a random value, or the true skill level plus a random error term. Our model also supports the definition of an arbitrary, but even number of competitors, and arbitrary number of competitive rounds and a fixed scaling constant K.

At each round of competition competitors are selected at random, two at a time, until all competitors are matched. Each competition is simulated as a Bernoulli random variable with win probabilities defined by (1.1) based on the true skill level. Based on the outcome each competitor's Elo ranking is updated based on equation (1.2) using their current Elo rating. In this approach the Elo rating is an estimate of the true skill level. Our overall research goal is to evaluate how the Elo ratings converge toward the true skill levels. To assess that we consider several metrics:

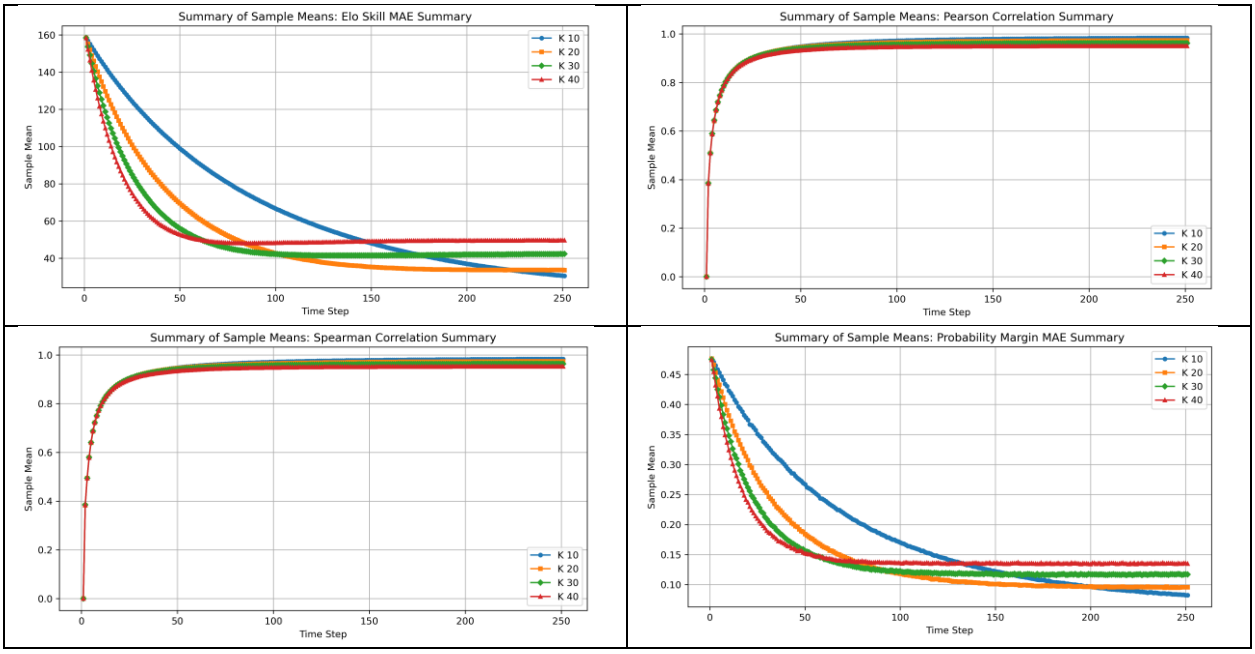
- Mean Absolute Rating Error: the mean of the absolute difference between the competitor's Elo rating and their true skill.
- Rating Correlation: the correlation between the competitor's Elo rating and their true skill, evaluated using either Pearson correlation coefficient, or the Spearman rank order correlation coefficient.
- Mean Absolute Probability Error: the mean of the absolute difference between the probability difference between competitors based on the Elo rating and based on the skill level.

Over time as competition plays out, we would expect to see competitor's Elo rankings converge toward their true value, so the two error metrics should decrease toward zero, and the correlation should increase toward one. If competitors are ranked in the correct order, regardless of an error in the absolute rating levels, the Spearman correlation should approach unity. If the relative difference between ratings is accurate, regardless of any bias in the actual scores, the Probability Error should trend toward zero. While the accuracy of the overall rating estimate is important, it is the ranking correlation and probability error metrics that are truly important measures. The key benefit of the Elo model is its ability to properly rank order competitors and properly assign win probabilities.

Experiment 1 – Large Population and Frequent Competition

Our first experiment deals with a large population and frequent competitions. This is the type of environment the Elo model was originally created for. It applies well to chess, Go, and other games that have large populations with frequent updates. All competitors are given an initial Elo rating of 1500. The skills for the competitors are set randomly following a normal distribution with a mean of 1500 and a standard deviation of 200. We simulate 250 rounds for each competitor and run the simulation for 500 iterations for each configuration. We run the model for different values of *K*, varying from 10 to 40. This simulation runs for about 30 minutes on a reasonably powerful desktop computer. Figure 4 presents the round based average values for each performance metric.

Figure 4- Key Performance Metrics for Large Population Experiment Varying K



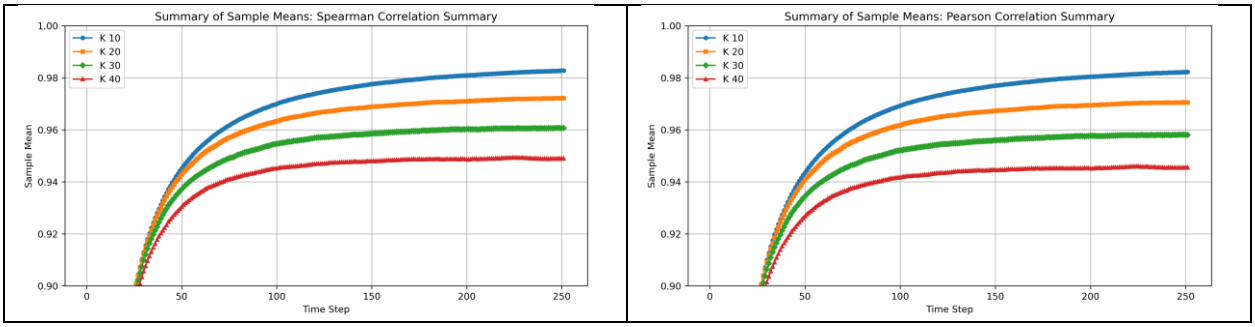
The first graph shows the MAE of the difference between competitor's skill level and their Elo rating. The error drops quickly, falling most rapidly for higher values of *K*. Simulations with higher *K* values settle in at a steady state error rating. The steady state error rate is inversely proportional to the value of *K*. A high *K* value converges faster, but to a larger error level. After 250 rounds the *K* equals 10 setting has still not reached a steady state error level.

The correlation coefficients between player's skill and Elo ratings increase rapidly. For the first 25 rounds or so there is little difference based on the *K* value. Figure 5 shows a zoom in of the correlations only showing values above 90%. In this figure we can see the long-term result of a higher K value, a slightly negative impact on the final level of correlation. With *K* equal to 10 the correlation exceeds 0.98, whereas with *K* equal 40 if settle in slightly below 95%. The difference in correlation values is negligible through the first 50 rounds, but by about 100 rounds the values diverge.

The final graph panel of Figure 4 shows the MAE of the probability differential. For obvious reasons this graph closely mirrors the MAE of the absolute skill difference. Higher K values

result in faster convergence, but higher steady state errors. The final error rates are quite high, with average error above 10% for K values of 20 or greater.

Figure 5 - Zoom In on Correlation Errors for Experiment



Overall, this experiment demonstrates that under the assumption of a large population with frequent competitions the Elo rankings generally do a good job of meeting the ranking evaluation criteria. The primary objective of achieving a reasonable ranking is achieved with a high degree of accuracy. Correlations over .9 are achieved rapidly and correlations in excess of .98 are possible. The experiment also indirectly validates the FIDE approach of assigning a high *K* value early to achieve general convergence and then lowering the *K* value for fine tuning.

Experiment 2 – Small Population, Infrequent Competition, Moderate Length Season

Experiment two involves a much smaller population and a moderate length season. Unlike experiment one this scenario has a finite length. For this experiment we set the number of competitors at 30 and the number of events at 82. This approximates the conditions of the NBA regular season.

Figure 6 shows how the key performance metrics evolve over the course of the season and Figure 7 zooms in on the correlation metrics. The general pattern is similar to experiment 1. Higher values of *K* lead to a faster reduction of the error on skill and the probability, but given the shorter season neither reaches a steady state. Correlations track each other rather closely through the first 20 games of the season, achieving correlation coefficients above .9 between games 20 and 30 of the season. Probability errors remain quite large for most of the season, and even at the end of the season remain between 10% and 20% on average.

Figure 6 - Key Performance Metrics for Small Population Moderate Length Season Experiment Varying K

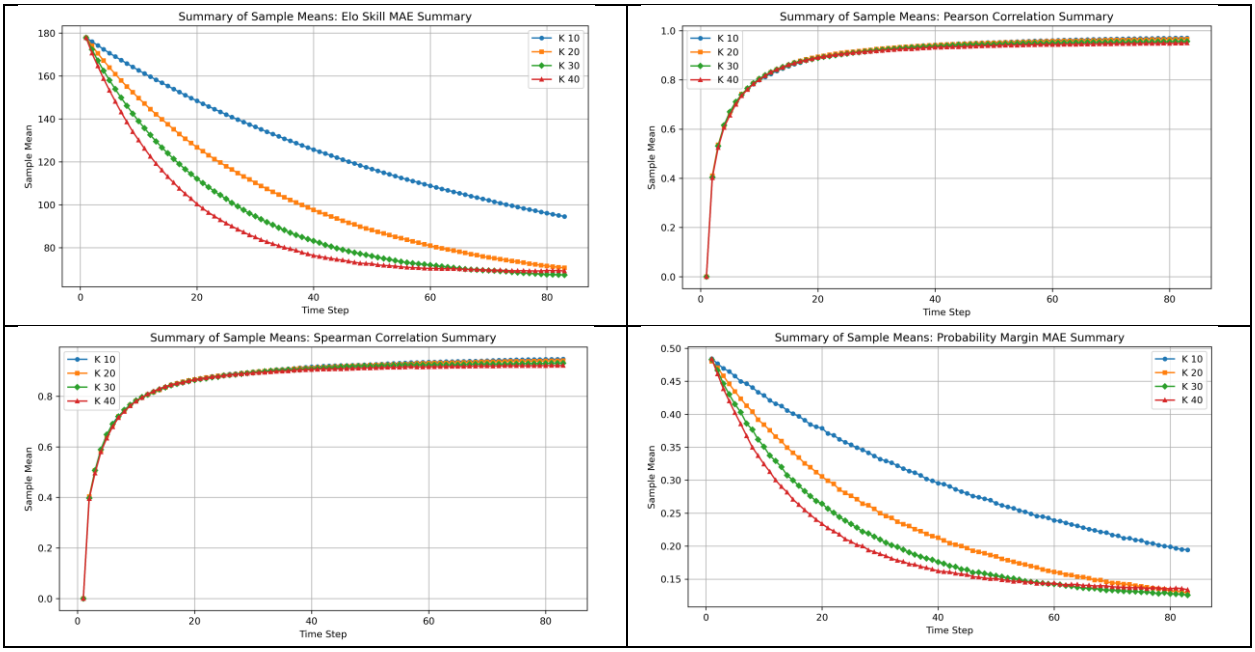
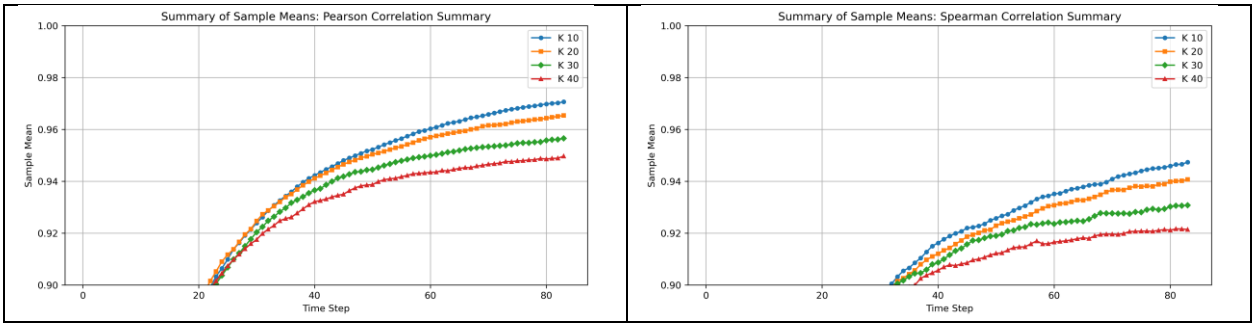


Figure 7 - Zoom In on Correlation Errors for Experiment 2.



Experiment 3 – Small Population, Infrequent Competition, Very Short Season

Experiment 3 is very similar to experiment 2, but the season is much shorter. Here we have 30 teams competing in 17 games. This scenario matches the NFL regular season. Relative to the length of the season the skill error converges slowly and remains high. For a *K* value of 20 the ending error is in excess of 100, ½ a standard deviation at the end of the season. Higher values of *K* again lead to faster error reduction in skill and probability error but have very little effect on the correlation metrics. The season is simply too short for the correlation metrics to diverge based on *K*. The value of the correlation coefficients is also relatively low given the short length of the season. Halfway through the season they are less than .8 and never reach a level of .9. The win probability differential errors also remain extremely high, in excess of 30% at the end of the season.

Figure 8 - Key Performance Metrics for Small Population Short Length Season Experiment Varying K

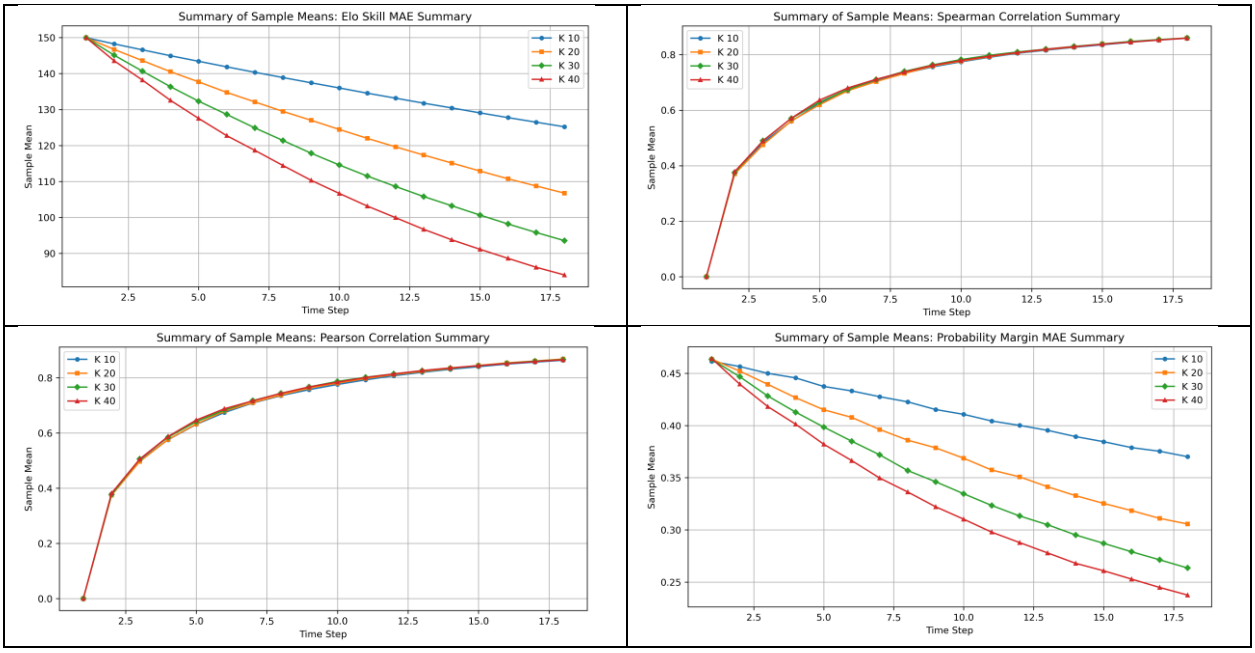
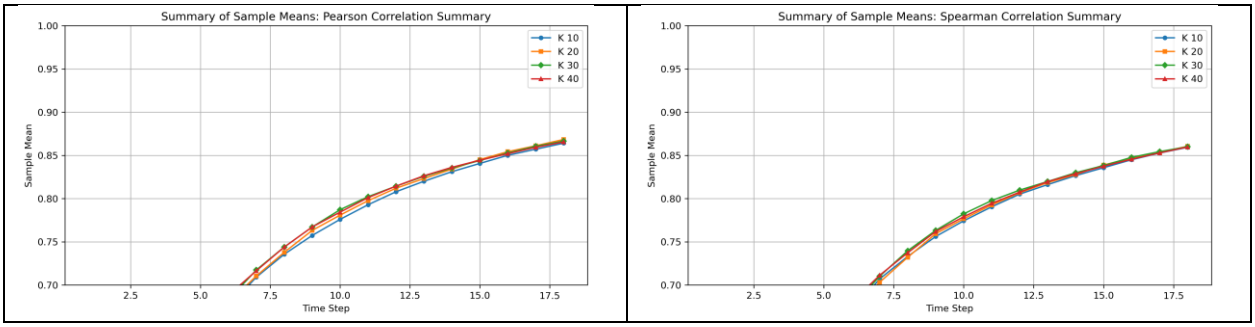


Figure 9 - Zoom In on Correlation Errors for Experiment 3



Experiment 4 – Small Population, Very Short Season, Preseason Rankings

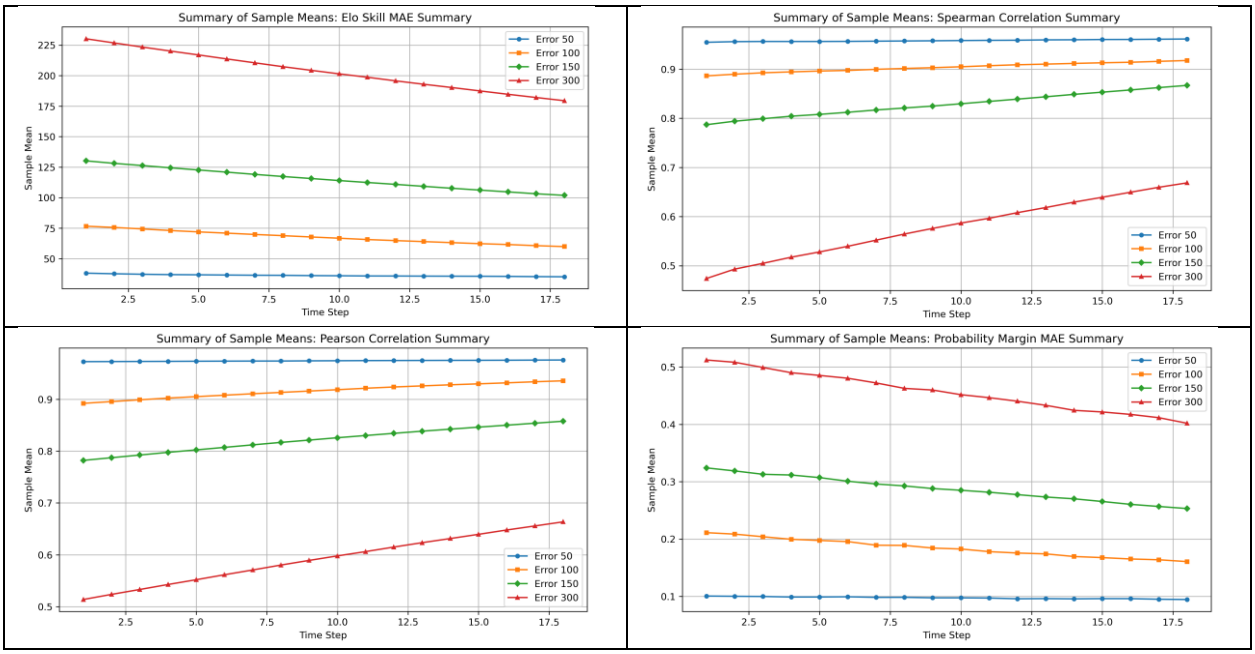
In the previous 3 experiments we initiated the experiment with a constant initial Elo rating. This is the system generally used with perpetual competitions, such as chess. While this works well for extended competitions, experiments 2 and 3 demonstrated that this leads to poor metrics through much of the season. In this experiment we examine the impact of a preseason rankling that is non-uniform. Specifically, we assume that the ranking at the beginning of the regular season is an unbiased estimate of the true skill level, with a normally distributed error term. This corresponds to the uncertainty introduced in the rankings during the offseason, for example changes in the coaching staff or in player personnel.

We evaluate scenarios with an error term standard deviation between 50 and 300. For this experiment we hold the value of *K* constant at 20 at execute a 17-game season as per the NFL. Figure 11 shows the standard metrics over the course of the season. Given that the initial rankings start out at least partially correlated with the true values, the rate of change for these metrics is much slower. Even in the case of an error term with a standard deviation of 300, the correlation coefficients start near 50%. But because the error is smaller, the points transferred

are generally lower, the high error correlations end lower than the uniform initial ranking given the short season. In the low error condition the initial rankings are relatively accurate, and little improvement is seen of the course of the season.

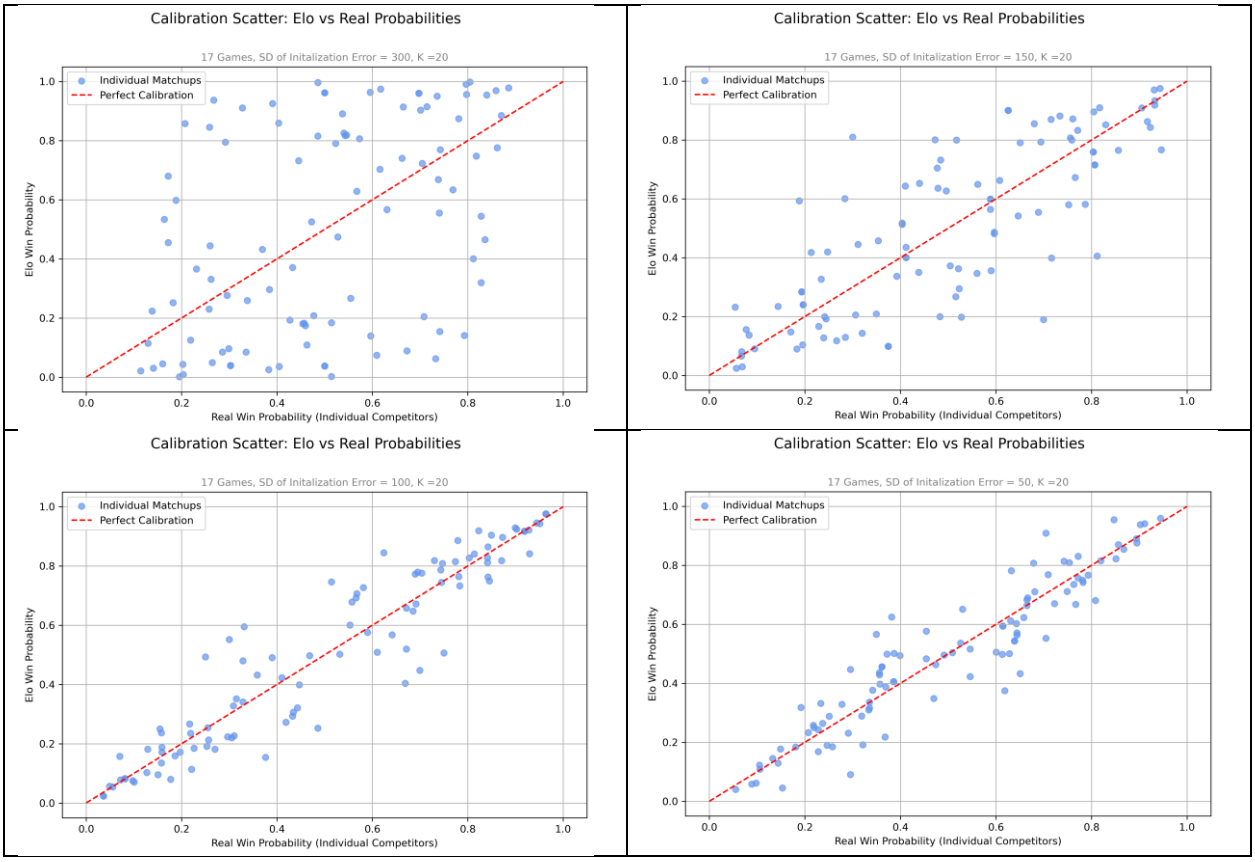
This experiment highlights the critical importance of an accurate pre-season ranking given a short season and justifies the focus placed on off-season adjustments by organizations such as FiveThirtyEight.com. With an accurate pre-season ranking, the ranking ordering of teams, as expressed by the correlation between skill and rating will achieve a high level early in the season. The absolute rating level error persists, but teams are effectively ordered.

Figure 10 - Key Performance Metrics for Small Population Short Length Season Experiment Varying Error



If we focus on the final goal of the Elo system, determining an accurate win probability for each competitor, the results are less encouraging. The probability margin MAE improves little and remains quite high, especially for the high error cases. To investigate this issue in more detail we generate a set of sample matchup after the last game of each season. We randomly sample 50 matchups from the league and compute the win probability for each team using the calculation based on the true skill level and the final Elo score. We developed a single sample for each error condition and present those results in Figure 11.

Figure 11 - Sample Team Probabilities at Seasons End



The highest error condition shows a very large error in win probabilities for most teams. The error reduces substantially as the initial ranking error decreases and with the best initial error condition, the Elo estimates are relatively closely aligned with the actual skill estimates. However, if we examine even the lowest initial error rate panel a clear pattern exists. Probabilities are most accurate at the extremes; teams that are in fact heavy favorites are estimated to be heavy favorites, and teams forecasted to be heavy underdogs are in fact heavy underdogs. The most significant errors occur when teams are more closely matched.

CONCLUSION

The Elo model is widely applied yet often criticized. A major issue with the model is that it assumes that each competitor has a stationary skill level. In practice the skill level is often non-stationary. The is especially true in team sports where injuries can dramatically change the skill level of a team during the season, and changes in personnel can change the skill level between seasons. For this analysis, we assume that issue away. We evaluate the model under the ideal condition where skills levels are fixed. Performance under these idealized conditions provides a bound on how well the model can perform under more realistic conditions. Future research should address how quickly the Elo rating can adjust to a new skill level or how well it can track a non-stationary skill level.

The Elo model has three distinct forecasting objectives; determine the competitor's absolute skill level, develop an ordinal ranking of competitors and define the relative win probabilities for an individual match up. With a sufficient number of updates, or an accurate initial setting, the

model achieves an accurate ordinal ranking as measured by the correlation between skill and rating. The absolute skill level is subject to a higher level of error that is persistent. The value of the constant *K* has a major impact on the rate of convergence and the final mean absolute error. Higher values of *K* lead to faster convergence but a higher mean absolute error.

The assignment of relative win probabilities is the most problematic of the three objectives. Even with a large number of matches the absolute probability error will settle it at between 10% and 15% for *K* values of 20 or more. In low frequency competition environments, such as football, the probability error start high and remains high through most of the season. Under the best case scenario we examined, where preseason ratings are set with an error term with a standard deviation of 50, the absolute error remains at about 10% the entire season. Even at the end of the season the error rate is the highest for evenly matched competitors.

In summary the basic Elo model works well in the environments it was originally designed for. Environments such as chess, where a large number of competitors with a large number of matches. Elo does best at rank ordering competitors and worst at estimating win probabilities. In short season competitions the model is weak during the early portion of the season, but improves later in the season. The quality of the forecast is highly dependent on the structured, and unstructured adjustments made during the off season to develop accurate pre-season ratings.

REFERENCES

Aldous, David. 2017. 'Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?', *Statistical Science*, 32: 616-29.

Angelini, Giovanni, Vincenzo Candila, and Luca De Angelis. 2021. 'Weighted Elo rating for tennis match predictions', *Eur. J. Oper. Res.*, 297: 120-32.

Badmintoncentral. 2010. 'Using Elo rating system to rank players within a club'. https://www.badmintoncentral.com/forums/index.php?threads/using-elo-rating-system-to-rank-players-within-a-club.91619/.

Bigsby, Kristina Gavin, and Jeffrey W. Ohlmann. 2017. 'Ranking and prediction of collegiate wrestling', *Journal of Systems Architecture*, 3: 1-19.

Carbone, Joel, Tony Corke, and Frank Moisiadis. 2016. 'The Rugby League Prediction Model: Using an Elo-Based Approach to Predict the Outcome of National Rugby League (NRL) Matches', *International Educational Scientific Research Journal*, 2.

cfbfastR. 2025. 'Get Elo historical rating data'. https://cfbfastr.sportsdataverse.org/reference/cfbd_ratings_elo.html.

Chen, Chang Heng, Joost N. Kok, and Willem J. Heiser. 2018. "Elo Rating System for UEFA Women's Euro 2017 The Predictive Power of Elo Ratings for the Performance of Teams and Players in the 2017 UEFA Women's Championship." In.

Elo, Arpad E. 1978. *The rating of chessplayers, past and present*.

Fargogate. 2025. 'Fargo Ratings'. https://fargorate.com/.

FIDE. 2025. 'Ratings Download'. https://ratings.fide.com/download_lists.phtml.

Goratings.org. 2025. 'Go Ratings'. https://www.goratings.org/en/.

Hvattum, Lars Magnus, and Halvard Arntzen. 2010. 'Using ELO ratings for match result prediction in association football', *International Journal of Forecasting*, 26: 460-70.

Kaggle. 2021a. 'FiveThirtyEight MLB Elo Dataset'. https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-mlb-elo-dataset/data.

———. 2021b. 'FiveThirtyEight NBA Elo Dataset'. https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-nba-elo-dataset.

———. 2021c. 'NFL ELO Ratings'. https://www.kaggle.com/datasets/dtrade84/nfl-elo-ratings.

Kovalchik, Stephanie A. 2016. 'Searching for the GOAT of tennis win prediction', *Journal of Quantitative Analysis in Sports*, 12: 127 - 38.

———. 2020. 'Extension of the Elo rating system to margin of victory', *International Journal of Forecasting*, 36: 1329-41.

Langholf, Fabian. 2018. 'The self-justifying Elo rating system', *arXiv: Classical Analysis and ODEs*.

Lehmann, Robert, and Klaus Wohlrabe. 2017. 'An Elo ranking for economics journals', *Economics Bulletin*, 37: 2282-91.

Morrison, Breanna. 2019. 'Comparing Elo, Glicko, IRT, and Bayesian IRT Statistical Models for Educational and Gaming Data'.

Nate Silver, Jay Boice, Neil Paine. 2023. 'How Our NFL Predictions Work'. https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

RacketInsight. 2025. 'USATT Ratings Explained'. https://racketinsight.com/table-tennis/usatt-ratings-explained/.

Robbins, Thomas. 2023. "Is the football Power Index a Good Bet?" In *Decision Sciences Institute (DSI) Annual Conference*.

ScrabbleAustralia. 2025. 'Tournament Ratings'. https://scrabble.org.au/tournaments/cgi-bin/rating.cgi.

Silver, Nate. 2014. 'Introducing NFL Elo Ratings'. https://fivethirtyeight.com/features/introducing-nfl-elo-ratings/.

StatCheck. 2025. 'Global Elo'. https://www.stat-check.com/elo.

Szczecinski, Leszek. 2020. 'G-Elo: generalization of the Elo algorithm by modeling the discretized margin of victory', *Journal of Quantitative Analysis in Sports*, 18: 1 - 14.

Tenkanen, Santeri. 2019. "Rating National Hockey League teams: the predictive power of Elo rating models in ice hockey." In.

tennisabstract.com. 2025a. 'Current Elo ratings for the ATP tour'. https://tennisabstract.com/reports/atp_elo_ratings.html.

———. 2025b. 'Current Elo ratings for the WTA tour'. https://tennisabstract.com/reports/wta_elo_ratings.html.

Vaughan Williams, Leighton, Chunping Liu, Lerato Dixon, et al. 2019. 'How well do Elo-based ratings predict professional tennis matches?', *Journal of Quantitative Analysis in Sports*, 17: 91 - 105.

Vaziri, Baback, Shaunak S. Dabadghao, Yuehwern Yih, et al. 2018. 'Properties of sports ranking methods', *Journal of the Operational Research Society*: 1-12.

Wikipedia. 2025. 'List of chess players by peak FIDE rating'. https://en.wikipedia.org/wiki/List_of_chess_players_by_peak_FIDE_rating.

Wolf, Stephana R. de, Maximilian Schmitt, Björn Schuller, et al. 2021. 'A football player rating system', *Journal of Sports Analytics*.