RJR Ratings System

World English-Language Scrabble® Players Association (WESPA)

Version 4.0

Russell Honeybun

Yong Jian Rong

8 October 2024

Table of Contents

Foreword and Acknowledgements

Karen Richards:

A rating system should be a "data" system. That means we can rely on it to inform us of facts. We should be able to play anyone knowing our chances of winning.

In 2019 I competed in Vadodara (India). Over 25 games, I met Madhav Gopal Kamath three times. Madhav had been competing successfully since the age of 6. He was now (aged 10) massively underrated by at least 400 points. The fact that I met him 3 times (the most of any of my opponents) suggests that his rating should have been close to mine. Because he won 2 of the 3 games, I suggest he was better than me. However, his rating was 220 below mine. I lost 11 points over the event. Had he been at his true rating, I would have gained rating points (I finished in the prize money). I joined this tournament knowing there was a risk I would have to play Madhav, or someone similarly underrated. However, many players would prefer not to compete in tournaments, where they know their ratings will be decimated unfairly, because their opponents are underrated?

We have desperately needed a rework of WESPA ratings for many years now (I contend, since 2006, when we started introducing significant numbers of new, rapidly improving players into the system.) My initial thought had been that we need to reinstall the previous "acceleration and feedback points". This allocated additional points to anyone who proved massively underrated, and also compensated any of their opponents. Jian Rong and Russell have worked extensively to make this system even better than I envisaged.

WESPA needs to ensure ratings are credible and reliable. Instigate a "data system", rather than an "inaccurate guesstimate".

Russell Honeybun:

Over the years I have been approached to review the rating system and asked if I could develop a better solution. The last time this occurred was the end of 2019. I was very interested in pursuing the ideas in a Taral paper I had been provided, but then it was 2020 and my focus switched to monitoring PPE levels across an area spanning 2.7 million km². We were locked behind a border for 18 months, protected from Alpha, Delta and the more aggressive variants of Omicron. Unable to enjoy interstate and international travel, I expend that time on non-Scrabble pursuits and was only playing one tournament a year.

2023 was a whirlwind, I blasted my expectations out of the water winning the South Australian champs, performing in the Aus champs, winning the West Australian champs and then placed 5th in the World's on Day 3 in Las Vegas, barely but still cashing. I carried that momentum through to WYSC Side tournament, again doing well. It was at that tournament, where the ratings for the Youth came out and I was asked a simple question by Karen Richards: "can you fix the youth ratings? The winners shouldn't be losing points, it's discouraging" I wasn't about to say no.

From there, I teamed up with Jian Rong who had TD-ed the Side Event and reviewed the current state of WESPA ratings. It became quickly apparent there was something amiss with the youth ratings, it was the entire system that needed fixing. The most obvious reason we could spot was that not all players were treated equally when starting out, and this disparate treatment then carried through once their rating was confirmed after just 50 games. Every player deserves the same chance to succeed. Simplicity first. Start

everyone on the same rating and see what happens. The results of such a basic change began an engine that saw me clocking past 2am on many nights, forgoing precious sleep and study to trouble shoot, build an analytics system and fitting algorithm and define transparent rules and concepts to test. (In fact I'm editing this document between days for the WA Championship!)

A rating system should be transparent, and robust with reproducible methodology and results. It should reward performance and consistency of effort, and it should be fair to the most number of players. I am proud to say that our efforts have produced just this.

Thanks to Chris Lipe for providing the SQL schema that allowed me to review the current state in such fine detail, without it I still might be trying to scrape together .TOUs files :P

Yong Jian Rong:

Being an enthusiast of probabilistic systems and applied probabilities, I have a natural interest for Elo rating systems. The motivation to start this project and revamp WESPA ratings came in 2017, when discussions with Chris Lipe introduced me to alternative rating systems such as Taral's Glicko 2 and Glickman's Glicko rating system in this paper.

I realised this project could have massive impact on the ratings of youth scrabble players, who have been underrated for the longest time since the beginning of WYSC in 2006. Thank you to Karen Richards for her unwavering support towards this revamp of the rating system. Thanks also go out to Russell Honeybun for providing and generating the important data visualizations that enabled new insights to be gained along the way.

The impact on player ratings arising from this paper will be felt across the player population. It will be felt more strongly for those who are working hard to improve their Scrabble skills and still stuck with 3-digit ratings from the past. With the new system, it is hoped that the WESPA ratings will gain the respect and recognition as a reliable source for referencing player's skill levels across nations.

References

1. Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. Journal of the Royal Statistical Society Series C: Applied Statistics, 48(3), 377–394. https://doi.org/10.1111/1467-9876.00159
2. Glickman, M. E. (2010). *The Glicko System*. The Glicko system. http://www.glicko.net/glicko/glicko.pdf

1. Introduction to Current WESPA Rating System

The WESPA rating system last went through a revision in 2011. In this new edition, a team comprising Karen Richards, Russell Honeybun, Yong Jian Rong with the support of Chris Lipe have come together to revamp the rating system and resolve the long-standing problems that youth players have faced in achieving an accurate rank/rating.

A rating system is a system that captures the relative skill levels of players in a zero-sum game. It is not a system of reward or punishment. Ratings work best in predicting win/loss outcomes when every player's rating accurately reflects their tournament performance based on their tournament records, and where the rating between two players and the relative expectation to win matches. A rating system should also treat the maximum number of players fairly, reward consistency and match the chronological record of a player's and nation's achievements. However, that is not the case currently.

Youth players are often the most adversely affected, with the vast majority starting out severely under-rated due to the legacy initial placement rating algorithm. That makes them prone to inherit the average rating of their opponents at youth tournaments. If the average rating of their first tournament is high, then that player is likely to receive a high rating. If the average rating is at the floor of the current system (300), then that player will again most likely inherit the average rating.

Their first tournament performance is so heavily weighted; it hinders the upward mobility of the players when they improve subsequently. Particularly, youth players improve in large increments in a relatively short period of time. If their win record matches their skill, they have to take rating points away from fellow youth and adult competitors to advance to higher categories in tournaments open to all ages.

Multiplied by many players in many tournaments, this deficit becomes stark and it becomes impossible to tell the different skill levels of players accurately. Players within the same country are all deflated to the same rating range or take points away from each other to advance. The next wave of youth competitors is then left with an even smaller pool of points to compete for. This effect is what we colloquially refer to as 'the youth problem', and this idea and how it was solved will be explained in a transparent detail with duplicable results in the treatise below.

A high-level overview of the problem definition, how it was tackled and the strategy the authors used is available in a sister document to this paper. A PowerPoint explainer used to present to the WESPA committee has also been made available. Please refer to the WESPA webpage containing this document for the appropriate links, or please reach out to the authors if you are unable to locate it.

This document presents a review of the incumbent WESPA rating system, followed by an introduction to the features of the Glicko rating system, the core model that the authors ultimately decided best fit the needs of not only the youth Scrabble players but of the entire world's competitive cohort. Finer details of the modification process to the rating system are available in the appendices

This revamp to the rating system aims to create an equitable starting point for all players while introducing the necessary mechanisms to provide accurate win expectations. For youth players, this change is most noticeable as the majority of players will be assigned ratings that are greatly improved upon when compared to the incumbent system. This is

a result of a conscious decision and analytical assessment by the authors as a lynchpin in creating fairer initial rating conditions for players at the beginning of their Scrabble journey.

Figure 1 below shows a preview of rating shift expected in the revised system, where the overwhelming majority of players will receive a boost to their current rating. On average, a player is expected to gain between 50 to 100 rating points as a result of this change. Youth players will experience larger rise in rating due to their currently low starting points.

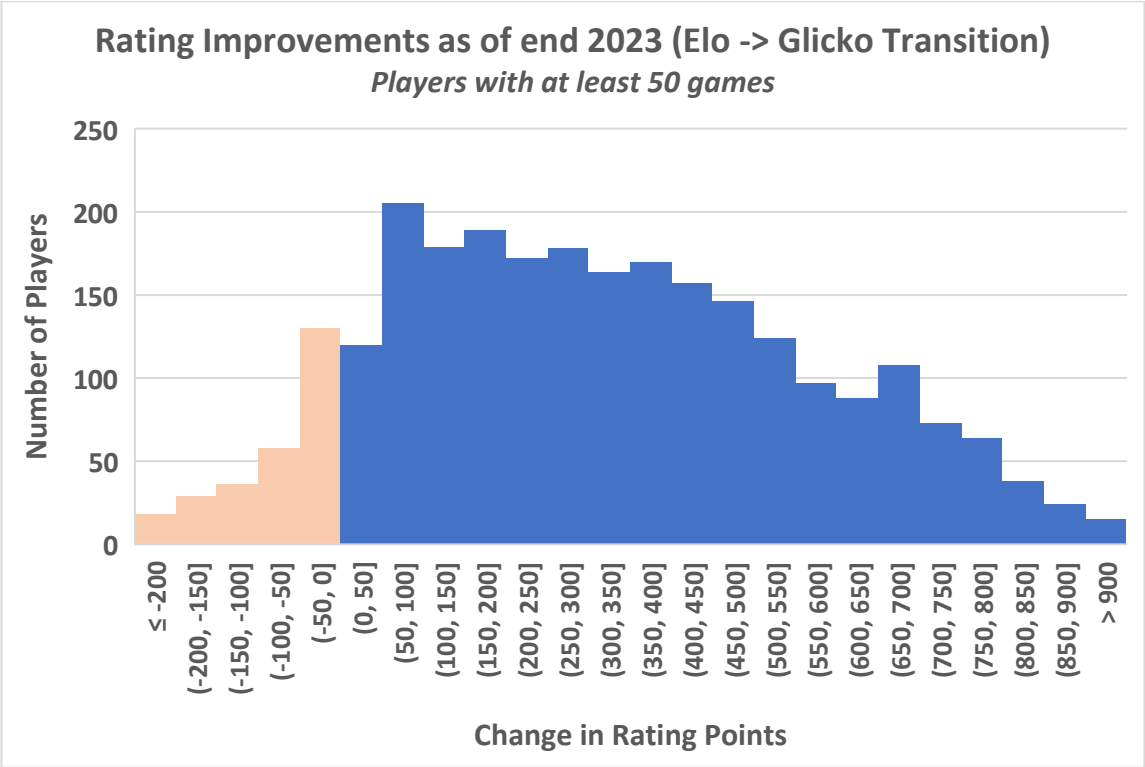


Figure 1 Population Statistics of Players receiving a change in rating due to transition across systems

1.1 Rating calculation for established players

The ratings system enables the ranking of players in 1-to-1 matches using a metric known as Elo rating. Using differences in rating, win/loss probabilities for each player-opponent pair can be determined. Counting each loss as a '0' and each win as a '1', the formula below can used to determine the win expectation[1] of each player-opponent pair. Rating is denoted using the shorthand *r, or sometimes R* depending on the example.

Expectation, $E_a = \frac{1}{1+e^{\frac{R_{player}-R_{opp}}{300}}}$

R_{player} = Rating of Player
R_{opp} = Rating of Their opponent
e = Euler's exponential (~2.718)

[1] Expectation or expected value refers to the average of possible values that a random variable can take. In tournaments, it refers to the win percentage of a player versus a specific opponent.

Note that in the above, '300' is referred to as constant $k = 300$. This number can differ across systems and determines how rapidly the win/loss expectation changes per point of rating difference. The denotation and understanding of k will be important later in understanding how k was varied under the new system.

Example 1

Under the current WESPA rating system, Player A rated 1700 plays against Player B rated 1500.

Expected probability of B winning = $\frac{1}{1+e^{\frac{1700-1500}{300}}} = 0.3392\ (33.92\%)$

Correspondingly, the probability of A winning can be obtained by switching the two rating values in the equation:

$$\frac{1}{1+e^{\frac{1500-1700}{300}}} = 0.6608\ (66.08\%)$$

In terms of odds, the odds of Player A beating Player B is approximately 2:1. More generally, any two players with a rating difference of 200 will have win odds ratio of about 2:1.

Currently, to convert probabilities into rating changes, a multiplier is used linearly such that rating change of one game = multiplier x result, then rating change across games are summed. Please refer to Example 2 for more granular details.

The incumbent system used 3 bands that changed as player rating increased or decreased. This is shown in Table 1.

Table 1 Rating multipliers for each band

Rating band	Multiplier
$r < 1800$	20
$1800 \leq r < 2000$	16
$r \geq 2000$	10

For two players within the same band (such as Players A and B), one player's gain is another player's loss. If Player A gains 10 points, Player B loses 10 points. There are no points gain or loss from the system using this method as it is zero-sum.

For two players from different rating bands, the rating points gain or lost by the two players can differ.

Example 2

Under the current WESPA rating system, Player A rated 1700 beats Player C rated 1900.

As the rating difference remains the same at 200 as per Example 1, the win probabilities remain the same at 0.3392 for Player A and 0.6608 for Player C.

As Player A wins and has a multiplier of 20, $win - expectation = 1 - 0.3392 = +0.6608$

$0.6608 \times 20 = +13.22$ rating points

As Player C lost and has a multiplier of 16, $win - expectation = 0 - 0.6608 = -0.6608$

$-0.6608 \times 16 = -10.57$ rating points

Only in situations where players meet opponents from different rating bands, will there be a difference in the amount of rating points gain/lost.

An iterative method is used instead if the player are completely new with no prior rating.

1.2 Problems Arising from use of the Incumbent Rating System

Under the incumbent system, a host of issues were identified, with the three most prominent being:

i. Inability to create new points
ii. Significant time lag in reflecting true skill levels
iii. Overweighted starting points

1. Inability to create new points

   Take the case where a player begins with the lowest possible rating of 300, at current rating multipliers of 20 points per game, the player would need to play 100 games against equal-rated opponents along the way (expectation = 0.5 win per opponent) to gain a theoretical maximum of 100 x 10 = +1000 rating points. This equates to taking away a total of 1000 rating points (to attain 1300 rating) from various opponents below the 1800 rating band.

   The problem entrenches itself when players with similarly low ratings are grouped together to match one another and compete for a small pool of points. The players are all of equal skill, and this prevents any upward mobility as the players as a whole gain a net of 0 points. Over time, this compounds and has resulted in many hundreds of players getting stuck at 3-digit rating levels despite having vastly improved skills compared to when they started.

   The problem can be understood better using various statistics compiled from existing WESPA rating data. In the Table below, countries with significant player populations were sampled.

Table 2 Partial Population Statistics

Country	Mode (Bin 50)	Average	Median	Number of Players
AUS	850	1032	1018	349
CAN	900	1358	1371	71
ENG	1350	1219	1240	213
HKG	300	787	749	64
IND	300	767	737	252
KEN	300	1080	1141	163
MYS	300	714	654	330
NGA	1850	1428	1490	289
PAK	300	690	583	199
SGP	1750	1241	1196	55
THA	300	860	789	271
USA	1250	1442	1449	144
Total	**300**	**1013**	**991**	**2400**

Players were grouped into bins of 50 rating points (rounded to nearest 50-point mark), and a count of players in each country with a given rating was performed. The results of this are shown in Figure 2.

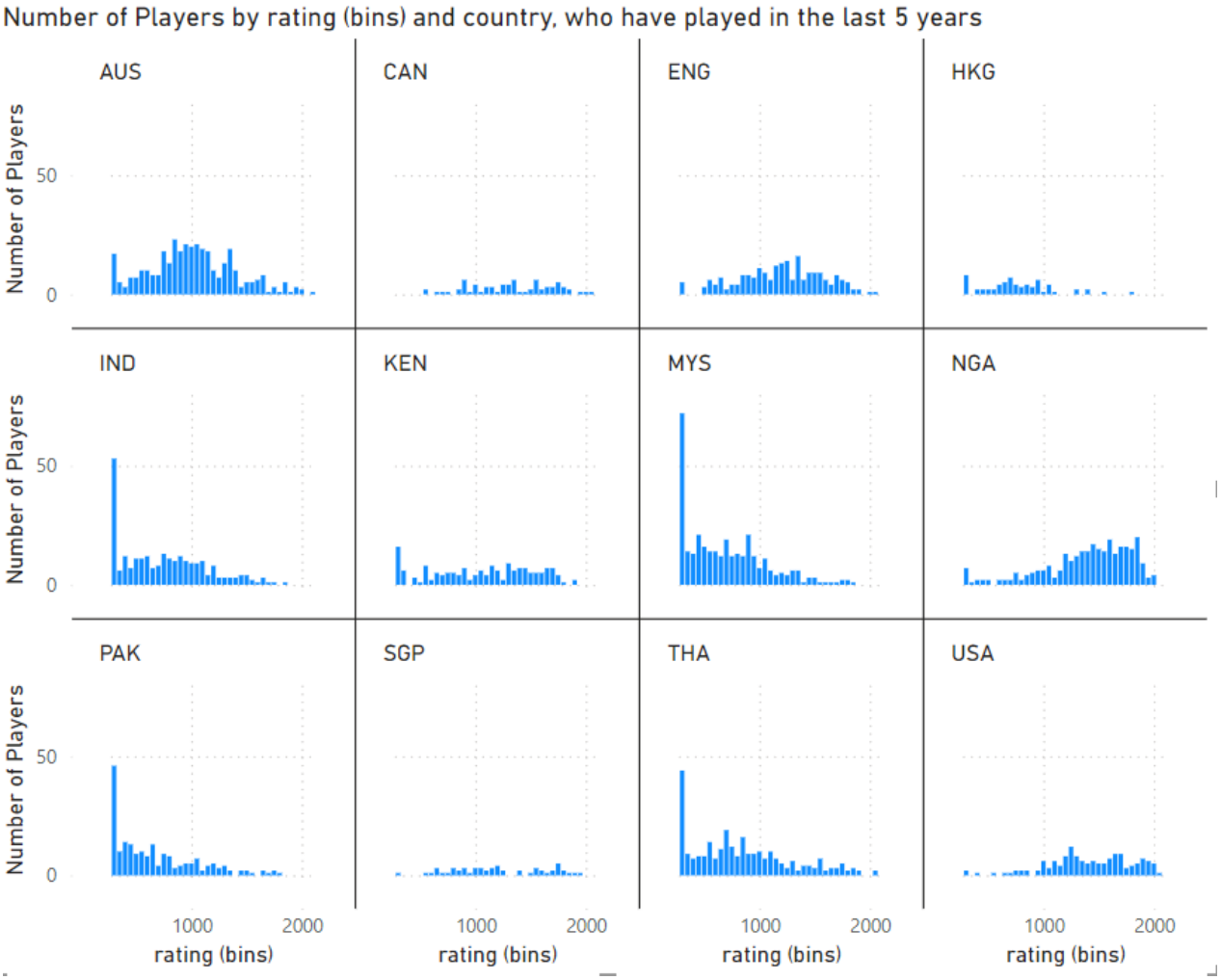


Figure 2 Rating distribution of players

In any population statistics, the expectation of any analyst is that they should be able to describe this population using the bell-curve or normal distribution. From the modal rating and histograms of various countries, it is evident that there is a significant pool of players stuck in the 300-rating zone. The shape of the distribution is far from being normally distributed.

Though the notion that the glut of low ratings is due to poor gameplay skills may veritably account for a small proportion, it is evident when comparing win % and scores across years and age brackets that the vast majority of these players are disadvantaged by their initial low rating and the low rating of their opponents. The problem then snowballs and entrenches itself as new players will encounter players with low ratings, and these players then fill the lowest-rated divisions, further preventing their access to higher-rated opponents.

2. Significant time lag in reflecting true skill levels

   The lack of WESPA rating for most local tournaments means that WESPA ratings often lag the player's true skill level. It is very common for players to only participate in a single WESPA tournament per year. In the incumbent systems, the magnitude of rating change remains constant regardless of how long the two opponents have been active or inactive. There is no accounting for the potential that one or both players may have improved tremendously. Hence there is a need to capture volatility or uncertainty in player performance after a period of inactivity.

3. Overweighted starting points

   Not every player starts with the same (dis)advantages. Primary differences in the opportunity to play other players of strength depend on factors of socioeconomic, cultural and geopolitically distinct nature to name but a few. The geopolitical aspect is described below, one completely out of the control of almost every player competing in their first WESPA tournament.

   In the incumbent system, a new player is assigned an initial rating relative to the average[2] of their opponent strength. This number is then iterated until convergence is achieved using their first-ever tournament's performance[3] to determine initial rating (also known as performance rating). Players who win few games may begin with a rating of 300, the current rating floor (lowest possible value). Players who win most of their games in their first tournament can begin with a high rating, which could be above the top-ranked and most seasoned player in the world. This iterative method assigns extremely high and undue weighting on the player's first tournament. Real-life case examples are explored in Example 3 and 4

   Example 3: Overtaking the world's top-rated player on 1st tournament

   As an established player within his country, Player C enters his first WESPA-rated event. The tournament had 15 games and Player C wins 11/14 games excluding 1 bye. Out of 11 wins, three were against opponents rated above 1900 and another four were against the Top 10 players in the world then. Player C finished the tournament with an initial rating of 2352 based on this calculation:

   Opponent's average rating = 1949.35
   1st iteration, rating = $1949.35 + \frac{400 \cdot 300}{172} \times \frac{2 \times 11 wins - 14 games}{14 games} = 2348.$

   After a few iterations[4], performance rating converges[5] at 2352. This was 80 rating points above the world number one player's rating then.

[2] If newcomer meets newcomer, that is excluded from the average.
[3] subject to lower and upper bound of 5% and 95% wins in the entire tournament
[4] see Example 1 for logistic equation used for iteration
[5] Convergence is reached when the player gains below 1 rating point for repeating the same tournament performance against his opponents.

In general, the formula used is:

$$Performance\ rating = Opp\ average + 1395 \cdot (\%\ won - \frac{1}{2})$$

This allows a maximum variation of ± 1325 at existing caps and floors of 95% and 5% wins, **regardless of the length of the tournament.** That means achieving identical % wins in a 1-day, 2-day or 3-day event will be regarded equally for newcomers. It is only for every player's first tournament that the percentage of games won is being used for computation, hence the fluctuations are extreme only for this first tournament.

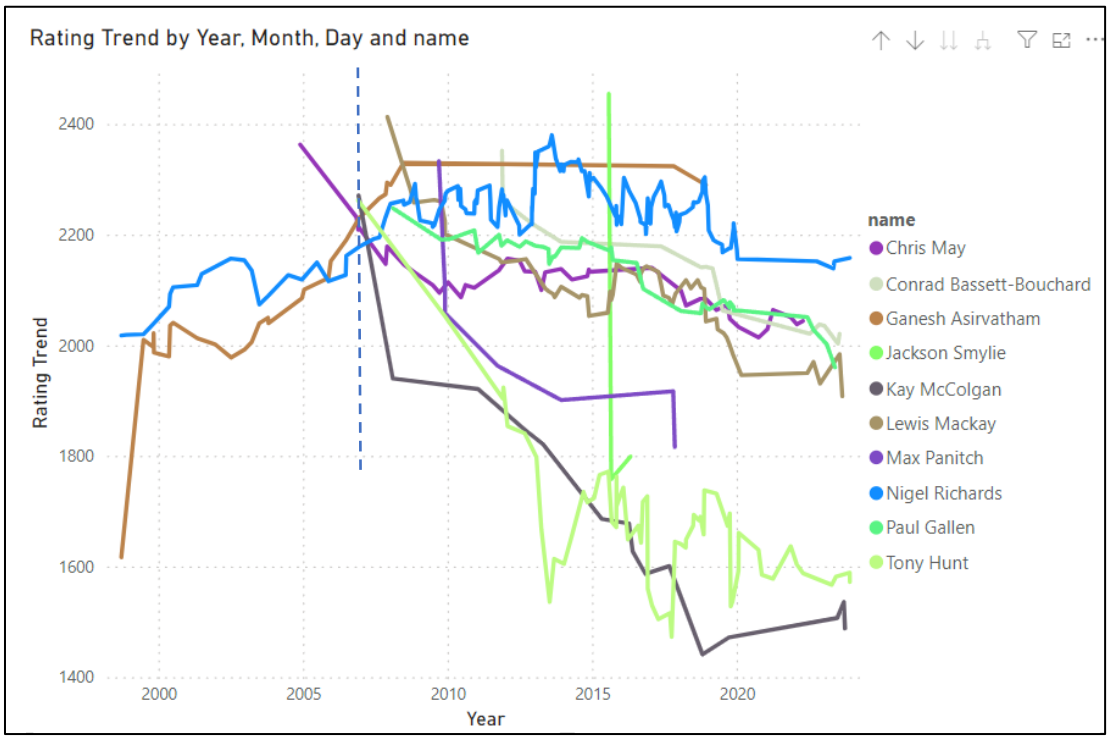


Figure 3 Rating Performance of Players with Top 10 Peak Initial Ratings

From 2005 to 2023, players have been affected adversely (to varying extents) by the initial rating iteration method, starting with high ratings and continuing on a descending trend ever since. Some extreme cases have been picked out as case studies above. The rating progression of Nigel Richards and A. Ganesh are shown for comparison.

The high sensitivity and high weightage of the player's initial rating relative to future ratings are issues to be addressed as a rating system should aim to capture the players' recent skill level after each tournament rather than being a slow-moving average. Also, inaccurate initial rating values that deviate beyond ± 100 rating points of a player's skill level should not be able to persist beyond 50 games if the algorithm is able to capture changes in skill levels.

Example 4: A bad 1st tournament with subsequent improvements

In contrast to Example 3, another extreme is Player A, a youth who started his WESPA rating at the rating floor of 300. In Player A's first tournament, the opponents both young and old were already under-rated by the current system at an average of 675 points. Hence with 6/27 wins, his rating was iterated until the floor was reached.

In three subsequent tournaments, he gained (240, 240, 156) rating points consecutively within a single month, a well-deserved massive gain that brought him closer to his true skill level. However, that was after he had taken away cumulatively over 600 rating points from all his opponents.

Due to the zero-sum nature of the current WESPA system, the number of points in circulation remains fixed and players with rapid improvements must remove corresponding number of points from their opponents. Opponents' average rating in the bottom division decrease over time while the improved player progresses to other divisions. That causes youth players to start out with lower initial ratings in later years.

1.3 Tenets of an improved rating system

With the problems identified, the authors experimented with various rating systems with an aim to achieve the following objectives:

1. Provide an equitable starting point for all players

   For newcomers to WESPA, rating changes should be proportional to the number of games won/lost and not by the percentage of wins/losses.
   The average of opponent's ratings will be considered but the value will be moderated towards 1500, a median value determined by experimentation.

2. Enable improved players' ratings to increase

   Players who have improved will be able to gain rating points at less expense to their opponents.
   That means the zero-sum scenario no longer holds and established players (>50 lifetime WESPA games) will gain or lose fewer rating points against a volatile yet improving player.

These two points above lead to the recommendation of the new system known as the **Glicko Rating System.**

2. Introduction to Glicko Rating System

The Glicko rating system was proposed by Dr. Mark E. Glickman of Harvard University in 1995. The mathematical details of the derivation can be found in a technical paper called "Parameter estimation in large dynamic paired comparison experiments"[6].

The Glicko system builds upon the Elo system by computing a rating along with a "ratings deviation" (RD), equivalent to a standard deviation used in statistics. It measures the uncertainty in a player's rating. A high pre-tournament RD indicates that a player has:

- just returned after a long period of inactivity; or
- only competed in a small number of tournament games; or
- performed erratically - winning lots of games against players with similar RD, but then also losing a lot of games to those same opponents

A low RD indicates that a player competes frequently and performs consistently against opponents who also compete frequently and perform consistently.

2.1 Comparisons with current WESPA system

Every effort has been made to align the results and outcome of the adapted Glicko system to rankings and results that players will be familiar with in the incumbent system. Variables have been finely tuned to present win/loss outcomes that resemble[7] those in the incumbent system, however the mechanics that drive those results are magnitudes different, as the calculation now relies on two primary variables (rating and RD) of the players and of the opponents. The incumbent system has a zero-sum property, meaning that for any rating gained by one player in a tournament, there will be an equivalent negative amount of points lost. This concept does not hold true in the new system.

There are extreme instances that may occur when the RD of all players in a tournament is the same (such as having 100% brand new players or small tournaments between very high-rated consistent players), producing equal magnitudes of rating gains and losses, but this will always be a highly unlikely exception and never a rule.

There are significant advantages to using a rating system that is not zero-sum. For higher-rated players losing to an up-and-coming newbie who is underrated while having a high RD, the higher-rated player's loss will be smaller than the gain of the new player. Conversely, as their RD is high, the new player can gain a larger number of points (sometimes up to two times as many) to catch up to the higher rated player.

When a player has been inactive for months or years, their pre-tournament RD calculated upon their return will lead to rating changes with larger magnitudes than if they had been playing frequently in WESPA-rated events. This presents an opportunity to 'correct' a player's rating in a transparent way based on their recent performance. The amount of correction and the inflation of RD has been finely tuned to the frequencies of play observed by both the average scrabble player, the youth contingent, and the elite cohort (the authors have identified that 'elite' in this situation is above 2000 ratings points, where

[6] published in the refereed statistics journal Applied Statistics (48, pp. 377–394), downloadable from http://www.glicko.net/research.html.
[7] Both make use of the logistic curve. WESPA k=300, Proposed k=250.

the RD can become the smallest and therefore, we also identify these players as highly-consistent performers.

2.2 New features in Glicko system

The following features are new/revised by the transition from a pure Elo system to a Glicko system:

- Steeper win expectation curve
- Newcomer ratings are initialised and calculated
- New interpretation: Rating & Rating deviation (RD)
- Rating deviation growth with time
- Opponent uncertainty function
- Opponent and Player Rating Deviation both affect Win Expectation

1. Revised Steepness to Win Expectation Curve

   k = 250 is used for the Glicko system while k = 300 is used for the WESPA system.

   This means each point of rating difference produces a slightly larger change in win expectations as shown by the steeper blue curve below. The grey line plots the difference in win expectation between the k=250 and k=300 curves. The maximum increase in win expectation is 4%, occurring at a rating difference of 450 to 500 rating points.

   4% translates to about 0.8 to 1 rating point per game, depending on opponent's rating deviation as explained in the next section.

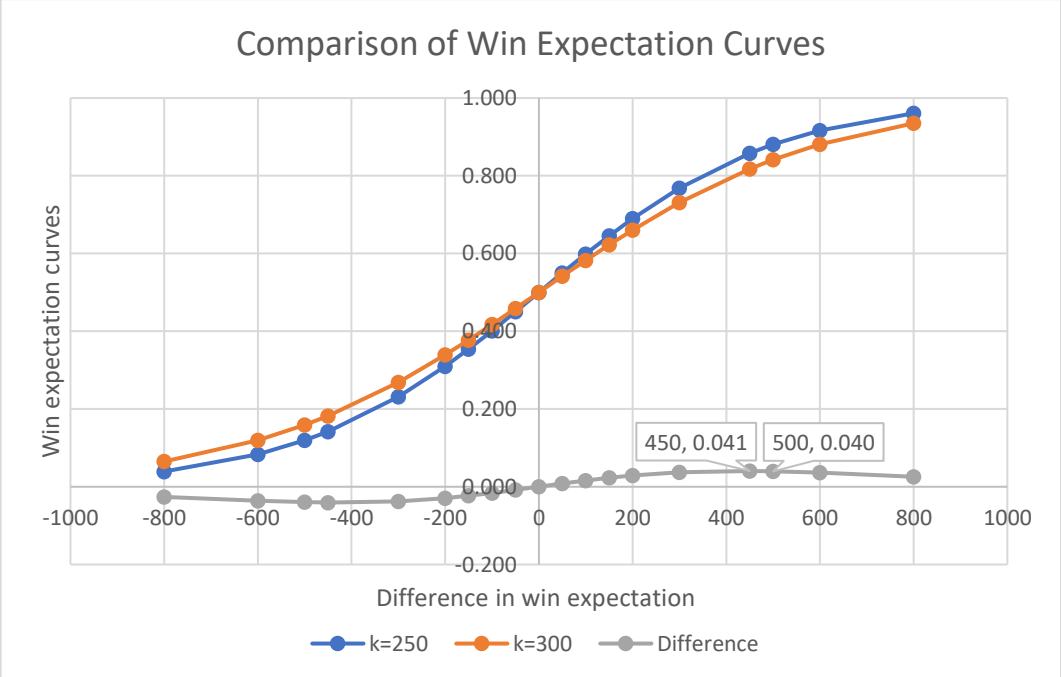
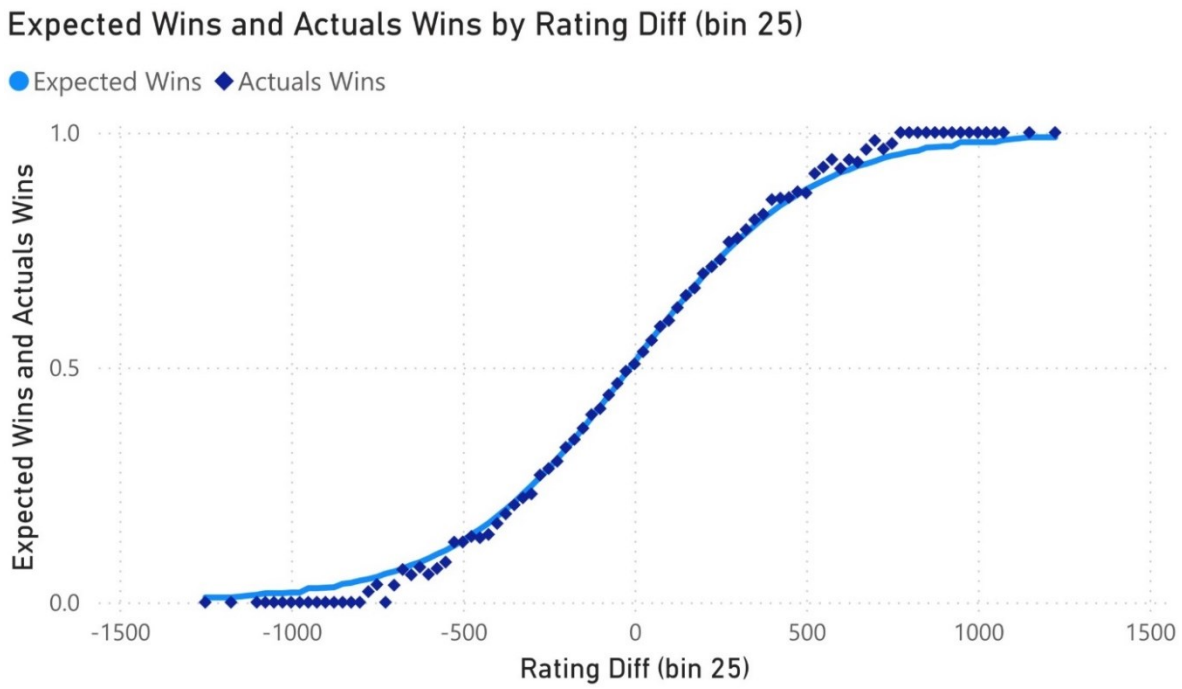


Figure 4 Win Expectation Curves (k=250 vs k=300)



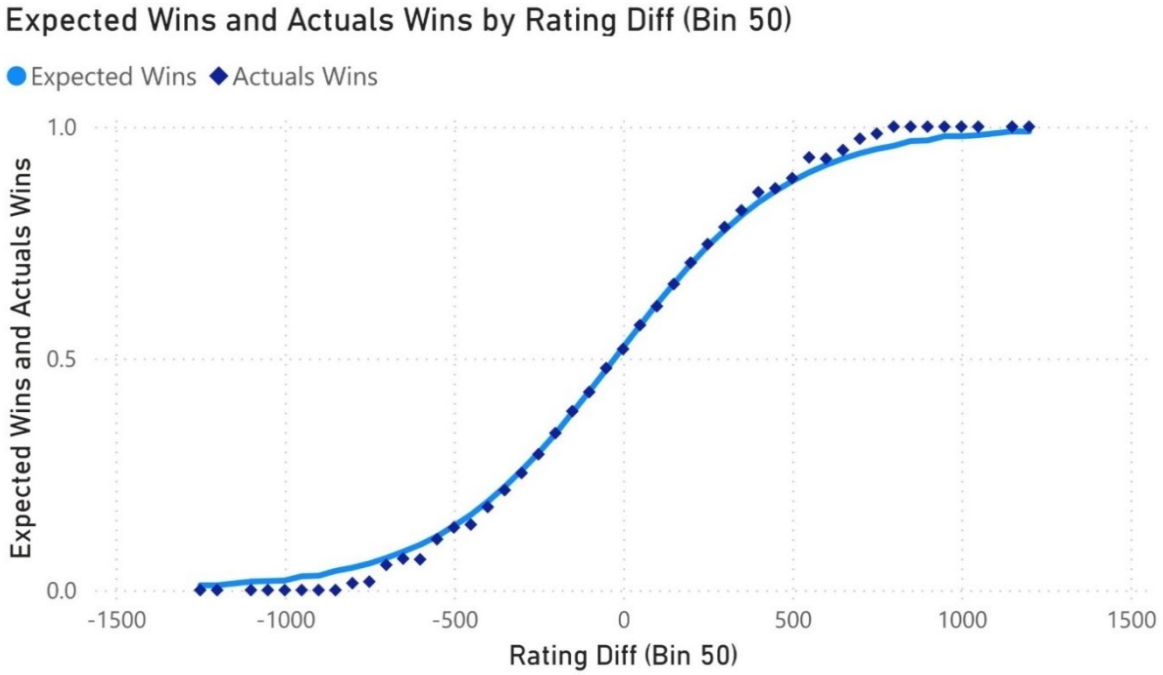


Figure 5 Expected Wins vs Actual Wins (aggregated by bins) (Glicko, k=250)

Comparing with actual win data, players were first grouped into bins of 25 and 50 rating points. Both graphs use the same data set of tournament games from 2015 to 2019.

Using k=250, Figure 5 above shows how actual wins tally with the win predictions using the improved Glicko rating system at the new k value. Dots denote actual win rates while the sigmoid curve is plotted using the average win expectations of players within that bin.

2. Rating Initialization

Upon entering a WESPA-rated event, every newcomer is assigned a pre-tournament rating which is then used to compute a rating change using the win expectation curve of the Glicko system.

It is computed by this formula[8]:

$$\frac{5 \times 1500 + Sum\ of\ rated\ opponents'\ ratings}{5 + Number\ of\ Games\ played}$$

The shorter the tournament, the closer the pre-tournament rating gets to 1500, the default starting value. This was implemented as past tournaments have had players playing fewer than 5 games and getting initialized either too high or too low. Through experimentation, weighing in 5 games was found to be optimal; it is sufficient for moderating short tournaments while not weighing heavily for longer events.

Be it due to players dropping out or the shortness of warm-up events, it would be unfair to selectively discard short tournament results as it sets a precedent in which players could "reset" their rating by dropping out on their first WESPA event.

Example 3a: Event with Highly-Rated Opponents

In a hypothetical tournament with all players rated at 2000, Player Y, a top-notch local player with no prior WESPA rating, plays 4 games and wins 2 games (50%) before dropping out. By winning 50% of his games against opponents of average rating 2000, his performance rating for this tournament is 2000 before adjustment.

His rating would be moderated as follows:

$$\frac{5 \times 1500 + 2000 \times 4}{5 + 4} = 1722$$

and thus the player is prevented from having a 2000 rating with only 4 games played.

Example 3b: Event with Low-Rated Opponents

In a 4-game warm-up event with all opponents rated at 1000, Player Z, a newcomer with no WESPA rating, plays 4 games and win 2 games (50%). By winning 50% of his games, his performance rating for this tournament is 1000 before adjustment.

His rating would be moderated as follows:

$$\frac{5 \times 1500 + 1000 \times 4}{5 + 4} = 1277$$

and thus the player is prevented from having a 1000 rating with only 4 games played.

[8] This formula seeks to prevent opponents in short tournaments (<6 games) from having excessive influence in the newcomer's initial rating.

3. Interpretation of Rating & Rating Deviation (RD)

One of the new and crucial features that distinguishes the Glicko rating system from the incumbent Elo system is the rating deviation (RD). RD measures the degree of uncertainty in the player's performance.

The rating and RD can be interpreted as a confidence interval. A player is expected to perform within $\pm 2\,RDs$ of his rating 95% of the time, assuming a normal distribution. For readers with a scientific or statistical background, this is equivalent to a 95% confidence interval.

Example 4: Seasoned Player vs New Player

Consider two players rated 1700. The seasoned player has RD = 70 while the newcomer has RD = 200.

The seasoned player is expected to perform at a skill level matching rating interval $[1563,1837]$[9] 95% of the time. In contrast, the newcomer has rating interval $[1308,2092]$ which predicts greater volatility in skill level.

We can map what the range of each player's skill looks like on a graph of frequency against rating. A higher peak on the blue line for the RD = 70 player suggests the player performs at the 1700-level more frequently (and other levels less frequently) compared to the RD = 200 player indicated by the green line.

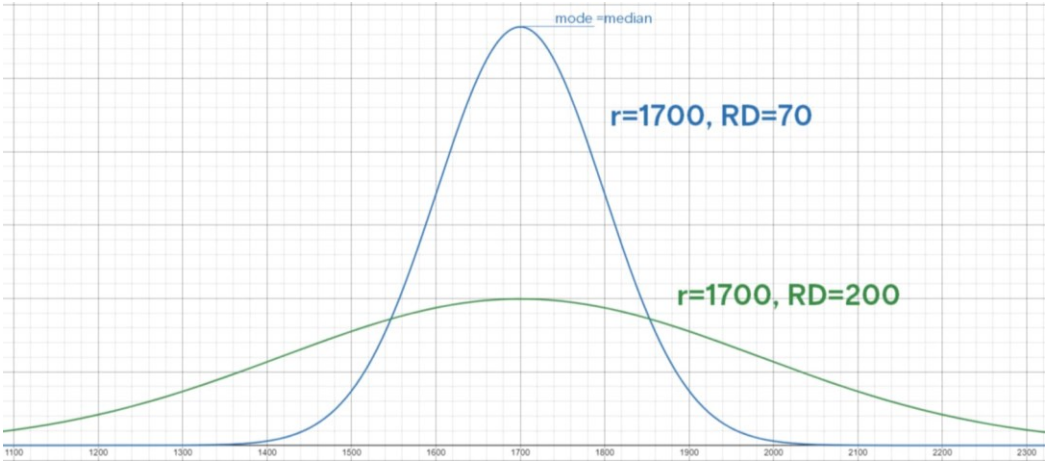


Figure 6 Rating Distribution comparison of consistent (RD=70, blue) and new player (RD=200, green)

The average rating change per game for the newer or less consistent player will be larger to help him achieve the correct rating range. Refer to Table 8 for calculations.

[9] 1700 – 1.96 x 70 = 1562.8 ; 1700 + 1.96 x 70 = 1837.2. '1.96' is the critical value for 95% confidence interval.

4. Rating Deviation Growth

Rating Deviation changes before and after each tournament. Upon processing a new tournament file, the participating players' RD will first grow according to c, a growth constant, and the number of weeks that the player has been inactive prior. The RD values of player and opponent are then used to decide the magnitude of rating change per game played. Note that a player's own RD has greater influence than his opponent's RD in affecting the player's rating.

These are the lower and upper bounds of the RD attainable for a player in each rating band. The upper bound is relevant to pre-tournament RD when a player returns from a long period of inactivity (between 3 and 4 years depending on their band). The lower bound is only relevant after the tournament conclusion.

Table 2 Minimum and maximum RDs for rating range

Rating band	RD_{min}	RD_{max}
No rating (newcomer)	-	300
$x < 1600$	75	200
$1600 \leq x < 1700$	70	175
$1700 \leq x < 1800$	65	150
$1800 \leq x < 1900$	60	125
$1900 \leq x < 2000$	55	100
≥ 2000	50	75

All new players begin with RD = 300 and an initial rating given by the formula in Section 2.2.2 Rating Initialization. As a rule of thumb, quadrupling the RD results in almost two times the typical rating change for identical game outcomes.

The increase from RD_{min} to RD_{max} for a player occurs over a period of 4 years (1461 days). The governing value of this factor, referred to as *c* in the formulae, was specifically chosen by the authors as it represents the typical interval between lexicon changes (Major historical updates for CSW were in 2007, 2012 ,2015 ,2019 and now 2025).

The equation governing this is:

$$RD = \min(\sqrt{RD_{old} + c^2 \cdot weeks}, RD_{max})$$

where c is a constant that controls the RD growth rate, adjusted to ensure RD_{max} is attained after 4 years. (*'weeks' = number of days/7)*

As it is tedious to calculate how RD grows with time, the table below provides an easy reference. Every 4 years has approximately 209 weeks.

Table 3 RD Growth for each rating band

		<1600		1600-1699		1700-1799		1800-1899		1900-1999		>=2000	
		Initial	Grown	Initial	Grown	Initial	Grown	Initial	Grown	Initial	Grown	Initial	Grown
Weeks inactive	Days	RDmin	RD	RDmin	RD	RDmin	RD	RDmin	RD	RDmin	RD	RDmin	RD
8	56		83.3		76.7		70.2		63.7		57.4		51.2
58	406		123.2		109.8		96.4		83.3		70.4		58.0
108	756		153.0		134.9		116.9		99.1		81.4		64.1
158	1106	75	177.9	70	156.1	65	134.4	60	112.7	55	91.1	50	69.7
208	1456		199.7		174.7		149.8		124.8		99.8		74.9
208.5	1460		199.9		174.9		149.9		124.9		99.9		74.9
208.71	1461		200.0		175.0		150.0		125.0		100.0		75.0
c		12.83		11.1		9.355		7.59		5.778		3.865	
RD Cap			200.0		175.0		150.0		125.0		100.0		75.0

As a rule of thumb, quadrupling the RD results in about two times the typical rating change for the same win/loss outcome against an opponent. Hence for a newcomer at RD = 300, the rating change per game would be twice as volatile compared to a seasoned player with RD = 75.

For players taking part in WESPA events once a year, their rating change would be 15 to 20% higher upon their return, than if two events were played back-to-back.

Example 5: Inactive Players Returns to Play

This example shows the actual output from a 2-day, 17-game tournament in late 2022.

Table 4 Sample results from a tournament

	17 games						
	Wins		Rating Points			RD	
Name	Exp	Act	Old	Change	New	Old	New
Player 1	13.4	14	2120	+12	2132	75	50
Player 2	9.2	12	1862	+44	1906	65	55
Player 3	10.3	11	1913	+13	1926	100	55
Player 4	8.6	11	1822	+70	1892	116	60
Player 5	7.2	11	1708	+125	1833	138	60
Player 6	5.8	10	1610	+194	1804	196	60

Among these players, Player 6 gains the most rating points due to a large difference of 4.2 wins (10 – 5.8), along with a large RD of 196. This results in an average of 46.7 points gained per game.

In contrast, Player 2 gains the least rating points due to the smaller RD of 65 and a smaller difference of 2.8 wins between expectation versus actual wins. This results in an average of 15.7 points gained per game.

Note that in both cases, the weight assigned to each of Player 2's and Player 6's opponent can vary such that the points gained or lost from each game differs. This is the uncertainty factor to be introduced next.

5. Opponent Uncertainty Factor, $g(RD)$

The opponent's uncertainty factor, referred to as $g(RD)$ in the Glicko system, is a factor which moderates the steepness of the ratings expectation curve. While the k-value defining the system does not change, win expectations can be affected by the $g(RD)$ term that is multiplied to it.

Adopting the definition from the Glicko paper, when computed at k=250, the function g can be simplified to:

$$g\left(RD_{player}, RD_{opp}\right) = \frac{1}{\sqrt{1 + 4.86 \times 10^{-6}\left(RD_{player}^2 + RD_{opp}^2\right)}}$$

For a game played against a non-provisional player (RD≤200), the output of function g is between 0.848 to 0.988.

Table 5 g-values for given RD values of player & opponent

Player RD	Opp RD	g
50	50	0.988
75	75	0.974
100	100	0.955
125	125	0.932
150	150	0.906
200	200	0.848

The output from g function is substituted into the logistic equation below, producing Table 6:

$$Expectation\ of\ the\ mode, Es = \frac{1}{1 + e^{g \cdot \frac{1500-1700}{250}}}$$

Table 6 Es at 200-point rating gap at different RDs

Player RD	Opp RD	g	Es
50	50	0.988	0.688
75	75	0.974	0.685
100	100	0.955	0.682
125	125	0.932	0.678
150	150	0.906	0.674
200	200	0.848	0.663

As observed, for a 200-point rating difference between two players, the value of Es changes by approximately 2.5% (0.688 - 0.663 = 0.025) as the RDs increase from 50 to 200 for both players. This small increment adds up to result in significant change over the course of 8-, 16-, 24- or 32-game tournaments.

6. Opponent and Player Rating Deviation both affect Win Expectation

Point 5 above and the Glicko paper give a conclusive breakdown of how a player's rating change and post-tournament RD are affected by their opponent's rating and RD across a multi-game tournament.

For players wanting to estimate their strength and likelihood of winning against another individual for a single game (sans tournament conditions) they may use $g(\mathrm{RD}_{\mathrm{player}}^2 + \mathrm{RD}_{\mathrm{opp}}^2)$. For different pairs of RD values, there may be the same Es outcome shown in Table 6. For instance, comparing squared sum RDs of $100^2 + 100^2$ and $120^2 + 74.83^2$, both produce the same sum of 20,000, resulting in the same moderation of expectation curve for the player and opponent involved.

However, note that the equation for computing rating changes places a higher weight on player's RD than those of the opponents. Summing up win expectations from single games and multiplying with a multiplier provides only an estimate.

The following graph plots win expectation at the mode against rating difference for three equal-RD scenarios: player and opponent having RD = 50, RD = 100 and RD = 200.

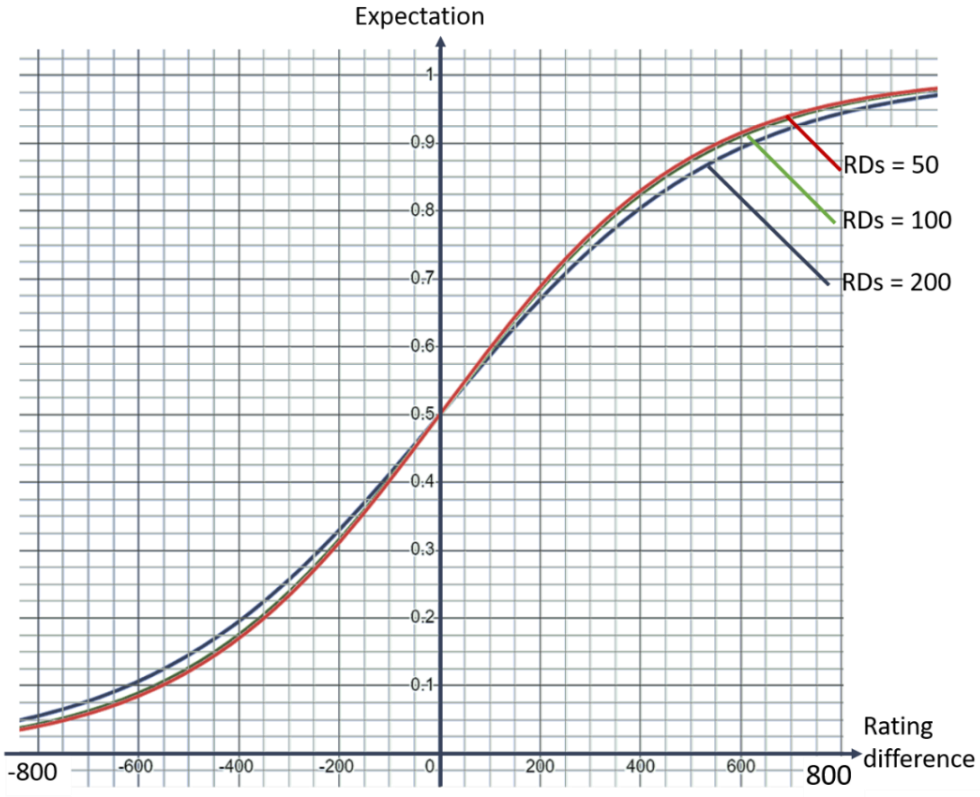


Figure 7 Effect of Higher RDs (lower g) on Es and Win Expectation

As RD of opponent increases, the win expectation of a single game gets less steep for the same gap in rating. That means winning against a high-RD opponent contributes towards a bigger overall rating change for the player, while losing leads to a slightly bigger loss in rating points. These differences are minute and adjust the win expectations by 2.5% at the maximum. At 20 rating points per game, this affects the rating change by up to 0.5 point per game. Note that rating changes

arising from these minute differences do not add up linearly as it considers the RDs of other opponents to produce an overall rating change.

To summarise, a gentler/less steep win expectation curve dependent on both player's RDs can contribute towards bigger rating changes, and this occurs due to either:

1. Higher uncertainty in the player's rating (bigger player RD)
2. Higher uncertainty in the opponent's rating (bigger opponent RD)

The effects of this are negligible for low-RD players (<150), but can help newcomers, typically with higher RD, gain rating points to reflect their true skill level faster after each tournament.

---

3. Impact Analysis of Glicko Rating System

The Glicko rating model was used to process tournament data from 1 January 2006 to 31 December 2023. To examine its impact, the ratings list produced by the Glicko and Elo systems were compared against each other. The rating data used below is as of 31 December 2023.

Figure 8 shows the rating changes (Glicko rating – Elo rating) among players who have played at least 50 WESPA-rated games since 1 January 2006.

Provisional players (<50 games) will take on their new Glicko rating as it is.

Out of 2583 players with ≥50 WESPA-rated games as of 31 Dec 2023:

- 2318 players will gain rating points, mostly ranging from +50 to +400.
- 265 players will have their ratings adjusted down to the new Glicko rating.

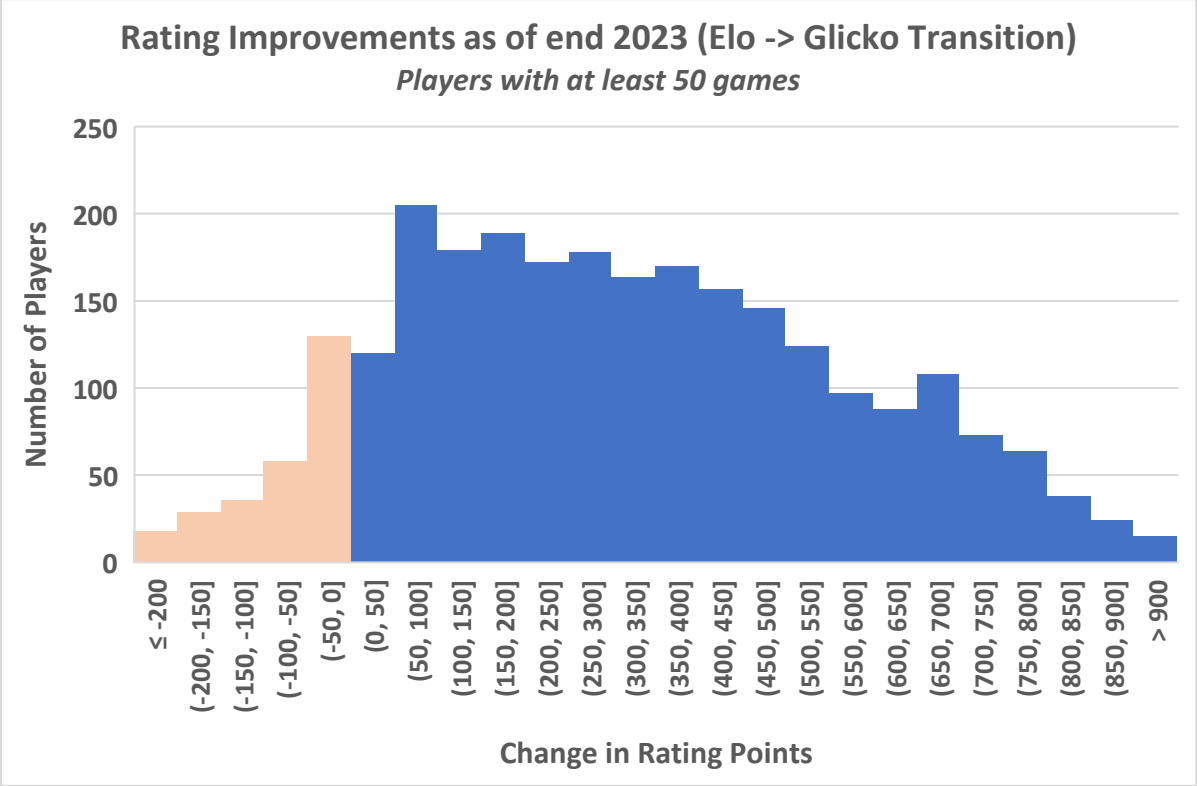


Figure 8 Changes in WESPA Rating (Elo → Glicko)

As a result of these changes, the revised player ratings are now normally distributed around a clearly identifiable mean.

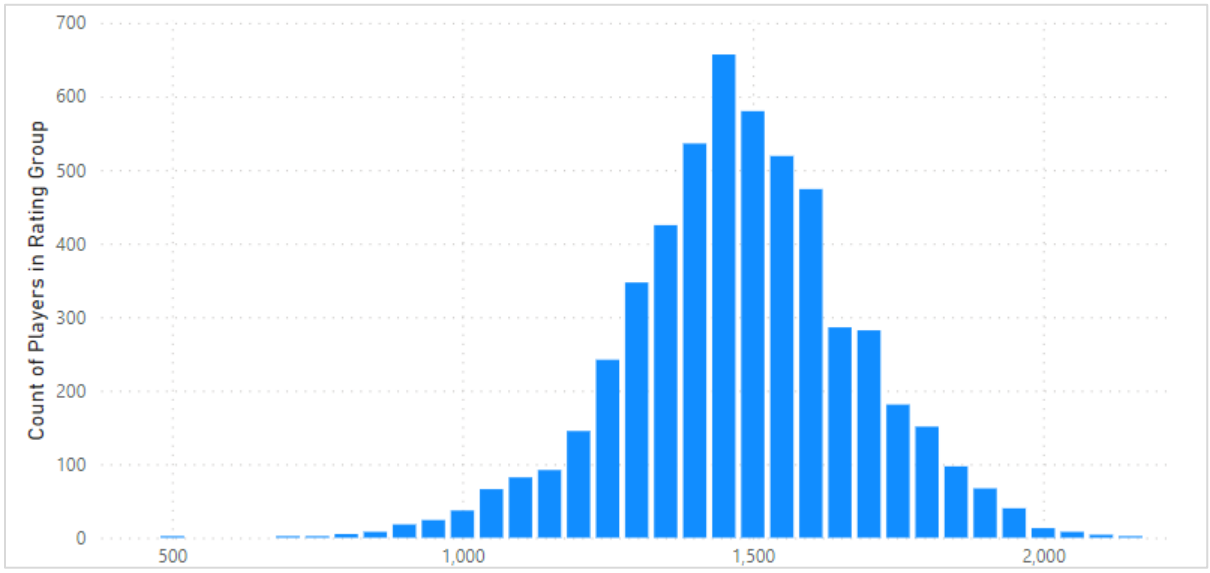


Figure 9 Player distribution across rating bands

Also, the number of matches played plotted against rating difference is normally distributed, with most players meeting opponents of similar rating and extreme rating differences occurring less frequently.

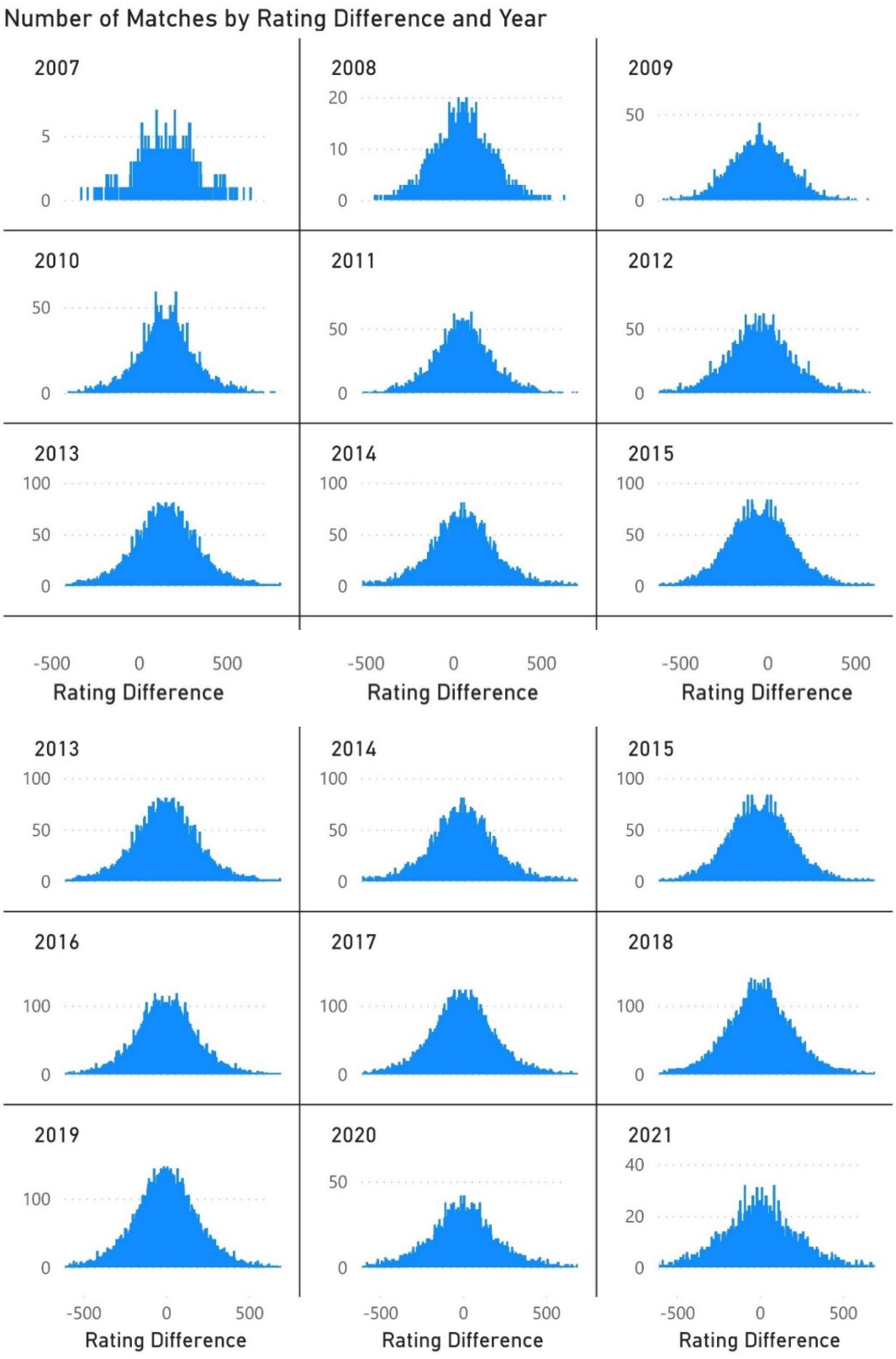


Figure 10 Distribution of player-opponent rating differences (by year)

The authors have provided an approximation to win expectation based on reasonable RD estimates of 70 for both player and opponent. Note the table below is an estimate for reference purposes. Calculations using the variables outlined in Section 2.2 are used to precisely calculate a player's rating, rating changes and RD.

Table 7 Rating Difference and 1v1 Win Expectation for Glicko system (k=250)

Rating Difference	**Expectation**	**Change per 10-point increase**	**Rule of thumb** *(Linear approximation)*
10	0.5098	0.98%	Win expectations: +1% per 10 points (10% per 100 points)
20	0.5195	0.98%	
30	0.5293	0.97%	
40	0.5390	0.97%	
50	0.5487	0.97%	
60	0.5584	0.97%	
70	0.5680	0.96%	
80	0.5775	0.96%	
90	0.5870	0.95%	
100	0.5965	0.94%	
150	0.6425	0.92%	+9% per 100 points
200	0.6860	0.87%	
250	0.7265	0.81%	+7.5% per 100 points
300	0.7636	0.74%	
350	0.7970	0.67%	+6% per 100 points
400	0.8268	0.60%	

Note: This expectation table assumes an RD of 70 for both players.

Table 8 Rating Change Estimate per game won/lost

	Maximum Points per game won/lost			
Pre-tournament RD	8-game tourney	16-game tourney	24-game tourney	32-game tourney
100	30	24	20.5	17.5
90	25.5	21	18	16
80	21	18	16	14
70	17	15	13.5	12
60	13	11.5	10.5	10
50	9	8.5	8	7.5

These point values are rounded to nearest 0.5.

Table 8 analyses the expected average rating change per game for tournaments of varying lengths. The calculated values assume that opponents encountered are identical in rating deviation (Opponent RD = Player RD = 70) for every game played.

Win = 1, Draw = 0.5, Loss = 0

Rating change in a game = (Outcome - Expectation) x Points per game

The two examples below demonstrate the use of Table 8.

Example 6: 8-game tournament with identical players rated 1700

In an 8-game tournament where all players have 1700 rating and pre-tournament RD of 70, any two players matched up should have an equal chance of winning/losing (50%). Hence over 8 games, the overall expectation for a player is about 4.0.

If a player wins 6 out of 8 games, estimated rating change = (6.0-4.0) x 17 = +34 points

If a player wins 2 out of 8 games, estimated rating change = (2.0-4.0) x 17 = -34 points

Example 7: 24-game tournament with identical players rated 1700

In a 24-game tournament where all players have 1700 rating and pre-tournament RD of 70, any two players matched up should have an equal chance of winning/losing (50%). Hence over 24 games, the overall expectation for a player is about 12.0.

If a player wins 18 out of 24 games, estimated rating change = (18.0-12.0) x 13.5 = +81 points

If a player wins 6 out of 24 games, estimated rating change = (6.0-12.0) x 13.5 = -81 points

4. Iterations leading to Glicko Rating System

A detailed chronology of model iterations leading to the decision and tested of the adapted Glicko system is provided here.

For every new model version, the results of every tournament file between 1 January 2006 and 31 December 2019 were processed and the results analysed. The initial iterations used this shortened time frame because of the unprecedented volatility that COVID-19 introduced. The remaining 4+ years was calculated for final results and the author's interest.

A primary goal of the analysis was that, after adequate time and matches, the expectation curve and actual win curve should have as low a weighted error as possible if the expectation curve reflects the underlying probabilities.

Several rating systems of different complexity were investigated to identify the pros and cons. These included:

Pre-Glicko Iterations

1. Rating by number of games played (simplest model, worst outcome, not modelled as it has no zero-sum properties)
2. Rating by number of games won/lost – 10 points per win and -10 points per loss, no other criterion. (Benchmark)
3. Ratings deconvoluted – where all existing WESPA ratings bands have the same multiplier of 20, and the rating bands are multiplied out by 20/multiplier
   (e.g. 1900 = 1800 + 20/16 x 100 = 1925 when deconvoluted)
4. Rating using linear expectations - each player's multiplier is determined by an equation that takes in their rating and produces a multiplier value between 5 to 40.
5. Rating using various sigmoid curve 'k's (Elo rating system) - optimising win expectations to match actual wins observed in tournament data

Post-Glicko Iterations

1. Adapting Glicko system following Glickman's paper – Max RD=350, k=210 to get initial data shape (Glicko benchmark). All players were equally uncertain and volatile in performance. RD had no lower bounds so it tended towards small values (<30) as players accumulated more games. This produced smaller rating changes as the years went by for each player.
2. Glicko system, with appropriate lower and upper RD limits identified and set. RD was mapped against rating, and a trend was found where average RDs 'settled' past the 1800 mark.
   Initial ratings algorithm was developed using the tried and tested Glicko methodology of initial rating = 1500, RD = 350.
   Past WESPA rating data were calculated using a limit on win possibility (like max 95%), so we implemented a change to the key algorithm to force a 95% maximum/5% minimum win prediction per game. Max RD was set to 300 and Min RD was set to 50.
   At this stage, the size of rating changes was also tuned based on what games between two RD=50 opponents should look like as we wanted the magnitude of rating changes to feel familiar to the current changes in ratings after a tournament.
   A concept of RD growth factor c was introduced. This brought the RD from

minimum to maximum RD values over a span of 5 years as suggested by Glickman. This was later fine-tuned to 4 years.

3. Poor results and further analysis led to refinement of the training dataset to players who had the following characteristics:
   i. played ≥50 games and RD<150
   ii. games had to have an opponent with the same characteristic

   At this stage, a variable q used in the calculation of pre-tournament RD (accounting for RD growth due to inactivity) was revised to q=1/k, improving on Glickman's $q = \frac{\ln 10}{400}$ that was intended for k = 400 and a logistic curve with a base of 10. $q = \frac{1}{k}$ suits our use case with base e and k = 250.
4. Later iterations showed that the win possibility limits of 5% and 95% came as a byproduct of high RD and would become unnecessary with smaller RD and more games played. Logistic curve least-square regression suggested removing limits on win possibility in every iteration to achieve smaller prediction errors from the regression best-fit model. Various 'k's were tested until its value settled at k=252 for RD 50. Upon including players of higher RD, the models became better at slightly lower values of k. Eventually, k=250 and 0%-to-100%-win possibilities (no limit) gave results with >99% accuracy of fit.

4.0 Binary Rating System (Pre-Glicko Iteration #2)

To show that there is a need to consider a player's actual performance with reference to expectations, the team first considered a model to show why the absence of win expectations is undesirable. All players were considered new as of 2006 and assigned an initial rating of 1500. Rating changes follow this rule:

$$Change\ in\ rating = \begin{matrix} +10\ if\ player\ wins \\ -10\ if\ player\ loses \end{matrix}$$

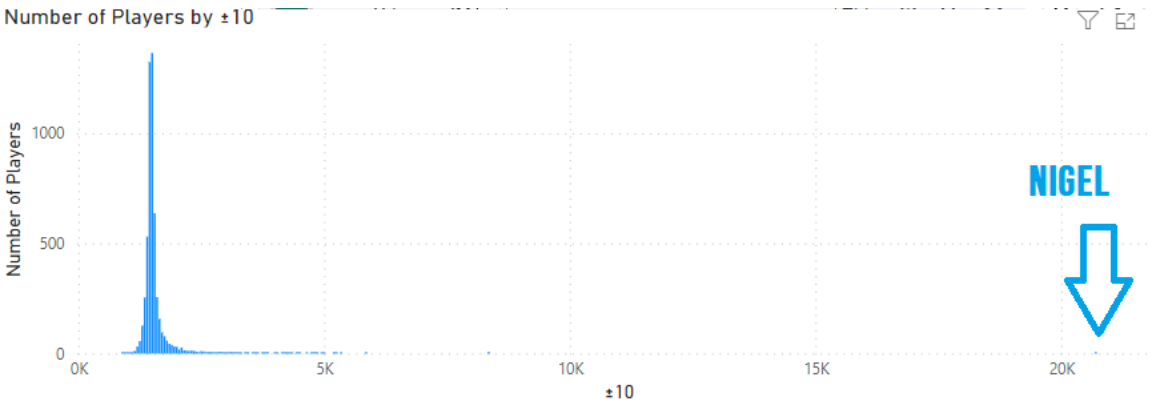


Figure 11 Player rating distribution of a Binary Rating System

The results as at end-2019 are expected – players who play and win more games will gain more points. If this model were to be adopted, it would result in farming, where veteran and active players gather most of the points and skill levels are irrelevant. Hence, there is a need to use a model that considers skill level captured in the rating itself.

4.1 Linear Rating System (Pre-Glicko Iteration #4)

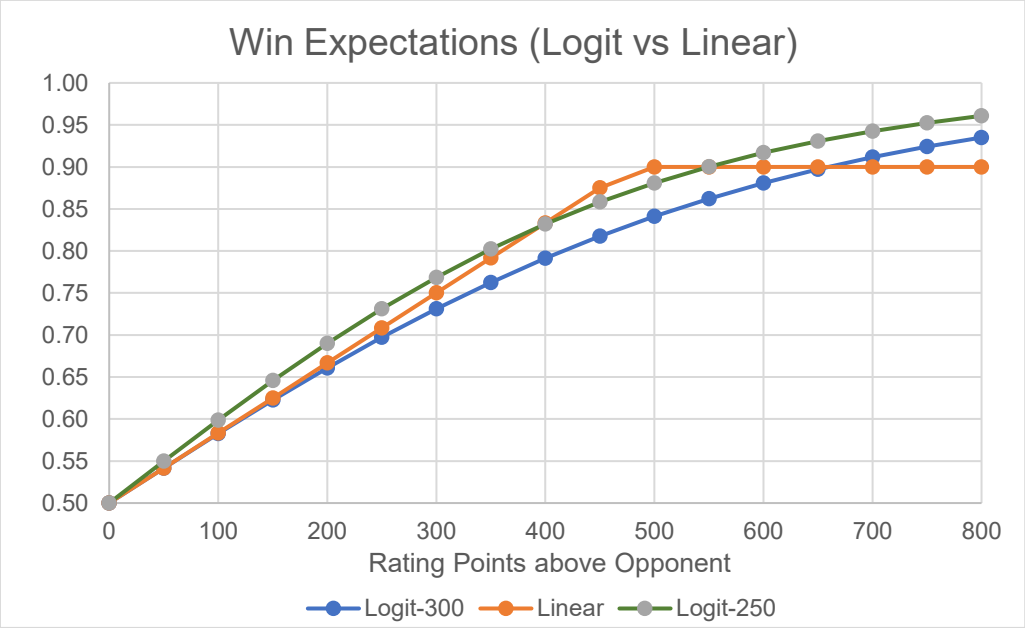


Figure 12 Win Expectations (Logistic & Linear)

This model was chosen as it had been used successfully in Australia for almost 20 years. Australia migrated from the ELO model similar to the incumbent WESPA system before choosing this one.

A linear expectation model was trialled with win expectations ranging between 10% and 90% following the rule: “Your percent chance of winning is fifty plus one twelfth of the rating difference.” Compared with logistic curves, Figure 12 shows how the linear expectation curve would look with caps at 90% and 10%. This cap ensures a minimum of 1 rating point gained/lost per game regardless of rating difference. The lower half of the expectation curve is not shown as win-loss expectations add up to 1.

On the next page, Figure 13 shows the win percentage of every player versus opponent result in year 2023 (97,000+ results) grouped into bins of 50 rating points. Based on two scenarios of capped win possibility (10% to 90%) versus no cap (0% to 100%), linear regressions were performed to examine whether the output equation matched the rule as coded.

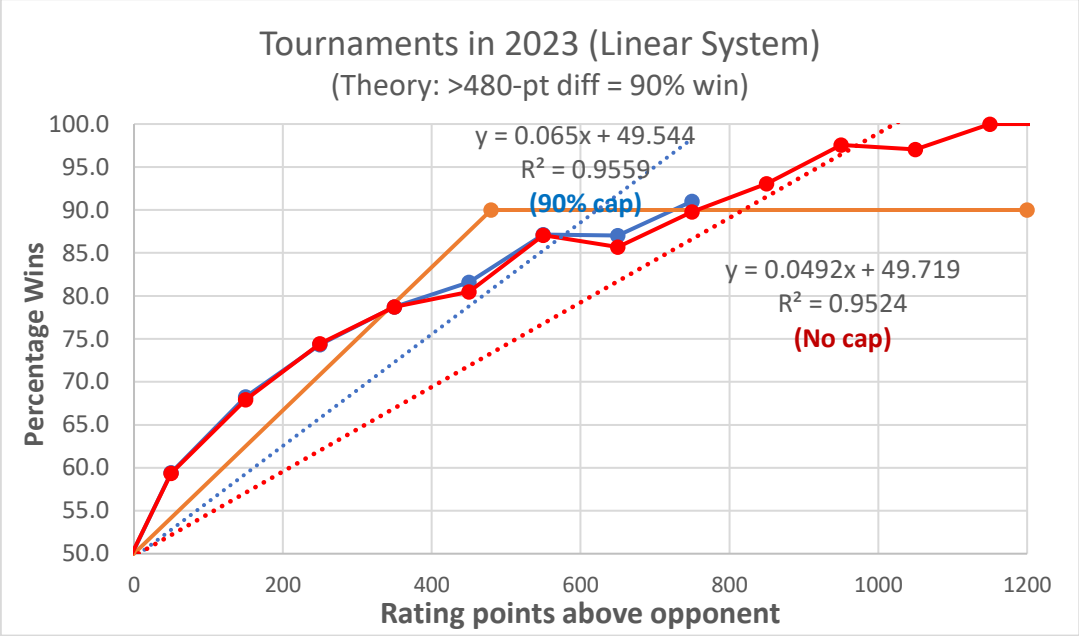


Figure 13 Linear Rating System (2023 Outcome)

In Figure 13, the orange line traces the linear expectations, while the red and blue lines reflect the actual win percentages grouped into bins of 50 rating points. Borderline cases (e.g. players rated at 1500 in one iteration but 1501 in another) may be re-grouped due to rating differences that propagate along the way, causing the blue and red lines to be different despite having identical tournament data.

For the no-cap iteration, the regression gave a win expectation gradient of 0.050, 1/20 the rating difference between player and opponent rather than the $1/12 \approx 0.083$ intended. For the capped iteration, the win expectation gradient of 0.065 was still >10% away from 0.083. For 50-point to 350-point rating differences, higher-rated players outperformed expectations against weaker opponents as both red and blue lines were above the orange line. For 350-point to 750-point rating differences, higher-rated players underperformed expectations, losing points on average.

This is a systemic problem as the linear expectation creates a separating equilibrium that tends to keep players at 300 to 400 rating points apart (with 350 as an approximate mid-point). Slightly higher-rated players (0 to 350 above opponent rating) overperform, taking rating points at the expense of opponents in the long term. Extremely high-rated players (450 to 750 above opponent rating) lose points to much lower-rated opponents in the long term[10]. This is despite the 90% cap that favours the higher-rated player by giving at least 1 full point per game won.

Consequently, one could exploit this separating equilibrium to flunk one tournament, dropping to a lower division to meet more opponents rated 0 to 350 points below them to farm a larger rating gain in a later tournament. Hence, this linear system promotes unstable tournament performance and thus unstable ratings.

[10] With enough up and coming youth players, players rated 450-750 above their opponents (with >85% exp win) do get defeated by lower-rated opponent, as tournament history has shown.

4.2 Logistics Rating System (Pre-Glicko Iteration #5)

As a result of analysis of the linear model outlined in 4.1, there was an observation that the actual wins appeared to approximate a curve rather than a straight line. The nature of the best-fit curve at this stage was unknown, so several sigmoid equation models was experimented by varying k-values. k = 230 is shown as an example.

$Es = \frac{1}{1+e^{-\frac{rating\ difference}{k}}}$ for k = 210, 220, 230, etc.

In setting the parameters, these changes (compared to current WESPA system) were considered:

1. Removing existing caps on overall tournament win percentage at 5% and 95% and removing the iterative process for newcomers.
    a. Players now have the full range of possible wins (0% to 100%) being used for rating calculation.
    b. Newcomers' ratings will be initialized to a pre-tournament rating, then changes are calculated only once for the tournament instead of iterating the same tournament many times until a converged post-tournament rating is found.

2. Adjusting k to curb excessive spreading of rating points from 300 to 2300 among the player population. Instead, a logistic curve with suitable steepness (determined by the k value) will be determined.
    a. There is little meaning in rating differences beyond 800 points as win expectations would be between 0.97 to 0.99, a range that creates less than 0.5 rating point difference.
    b. Rating ceiling and floor of 800 and 2400 are imposed, with the rationale that a median 1600-rated player will not meet opponents with rating differences of more than 800 points apart.
    c. Through making comparisons with legacy system (k=172) and current WESPA system (k=300), an initial Elo system (k=210) was trialled before increasing it towards k=230 in steps of 10.

Before interpreting results, it was first confirmed that equilibrium had been reached for a k = 230 logistic system. The rating outcomes of the k=230 trial was compared against other 'k's to confirm that deviations have been minimised.

Table 9 Win predictions fitted on other logistic 'k's

k	Weighted Average Deviation
210	3.04%
220	3.02%
225	3.01%
230	2.99%
235	3.06%
240	3.14%
250	3.28%

Table 9 confirms that the data for k = 230 are for a player population which has reached equilibrium, with win outcomes matching the k = 230 expectation curve most closely as compared to other k values. Calculation details of the deviations are shown in Table 10.

Table 10 Logistics k=230 Predictions vs Actual Wins

Rating points above opponent	Mid-point	Expect Win %	Actual Win %	Deviation	Weight
$0 \leq x < 50$	25	54.58%	52.71%	1.87%	5
$50 \leq x < 100$	75	63.34%	58.08%	5.26%	4
$100 \leq x < 200$	150	68.06%	65.75%	2.31%	3
$200 \leq x < 300$	250	74.72%	74.78%	0.07%	2
$300 \leq x < 400$	350	79.32%	82.08%	2.76%	1
$400 \leq x < 500$	450	82.43%	87.62%	5.18%	1
$500 \leq x < 600$	550	86.16%	91.62%	5.46%	1

Weights are assigned to the deviations to reflect the relative frequency of such games occurring. Heavier weights for 0 to 100-point differences emphasises the need for higher predictive accuracy for close match-ups. Using the weighted method above, weighted average error term = 2.99% for the k = 230 system.

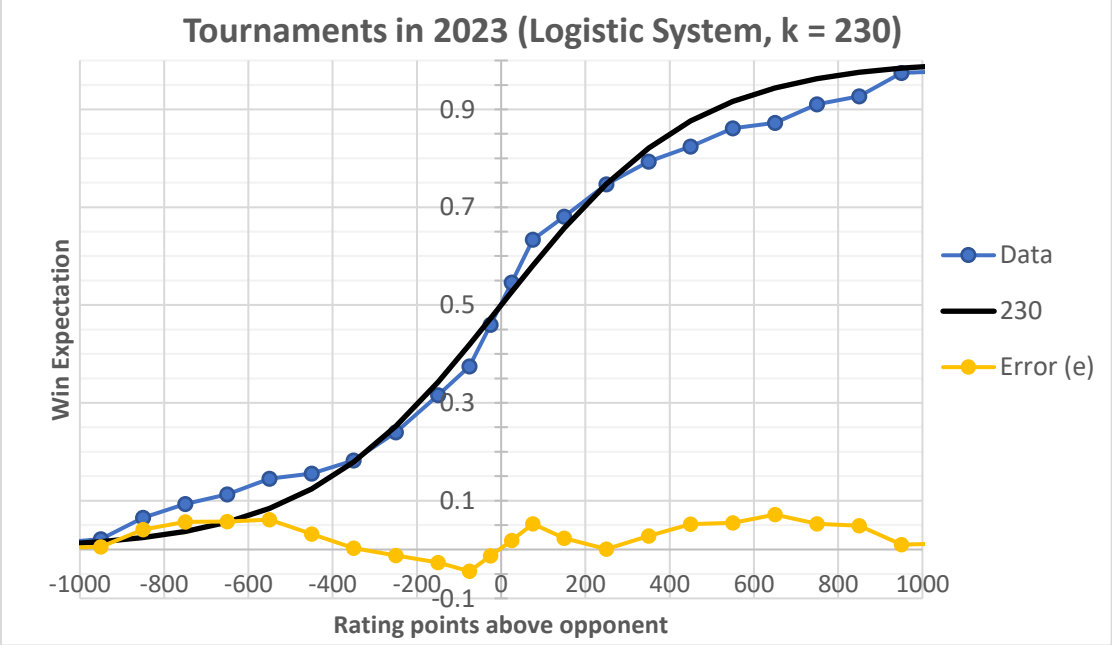


Figure 14 Logistic Equation System (k=230)

Looking at Figure 14, for match-ups involving players rated 0 to 200 rating points above their opponents, higher-rated players still overperformed their expectation as shown by the blue

line going above the black line. This would form a separating equilibrium where players' ratings tend to spread towards a difference of 200. Nevertheless, there is an improvement as the range of overperformance only extends up to 200 instead of the 350 reported in the linear system.

4.3 Glicko Rating System (Glicko Iteration #3 and #4)

In the next stage, a regression solver was used to speed up the solving process for the optimal k, with the application of g function to moderate the expectation curve. In preparing the formula, a change was made to term q, a constant within $g(RD)$, as the author assumed k = 400 and used a base of 10 in his calculations of win expectations.

On Page 3 of Glickman's paper:

$$q = \frac{\ln 10}{400} = 0.0057565$$

$$g(\mathrm{RD}) = \frac{1}{\sqrt{1 + 3q^2(\mathrm{RD}^2)/\pi^2}}$$

In the latest Glicko results, $q = \frac{\ln e}{k} = \frac{1}{250} = 0.004$ to match k=250 and the base e used in the baseline sigmoid curve equation. Consequently, the smaller q causes the moderation and flattening of expectation curves (as RD increases) to take place at a more gradual rate. Adjusting the constant q according to the k-value and base-e proved to be crucial in reducing prediction errors to below 1%.

Through observation of rating progression, it was observed that this settled after 3-4 years, and the highest quality and most relevant data occurring in the most recent time period, hence the regression solver refined its view to consider Glicko rating data only from 1 January 2015 to 31 December 2019. The Glicko rating system is sensitive to player inactivity, and as previously mentioned, ratings from years 2020-2022 were considered too volatile as a result of the effect COVID-19 had on the frequency of tournaments played across the globe.

Match-ups between players and opponents of identical rating differences (e.g. games with players having 300-point rating difference) were grouped to compute the average win rate for that rating-difference band. Some win rates were observed to be skewed or far different from expectation: this is due to small sample sizes at those extreme ranges. It should be noted in the comparative linear model investigation, a difference of 540 was considered to be the maximum. There was no such artificial limit imposed on the test data as the authors wanted an honest answer from the regression solver.

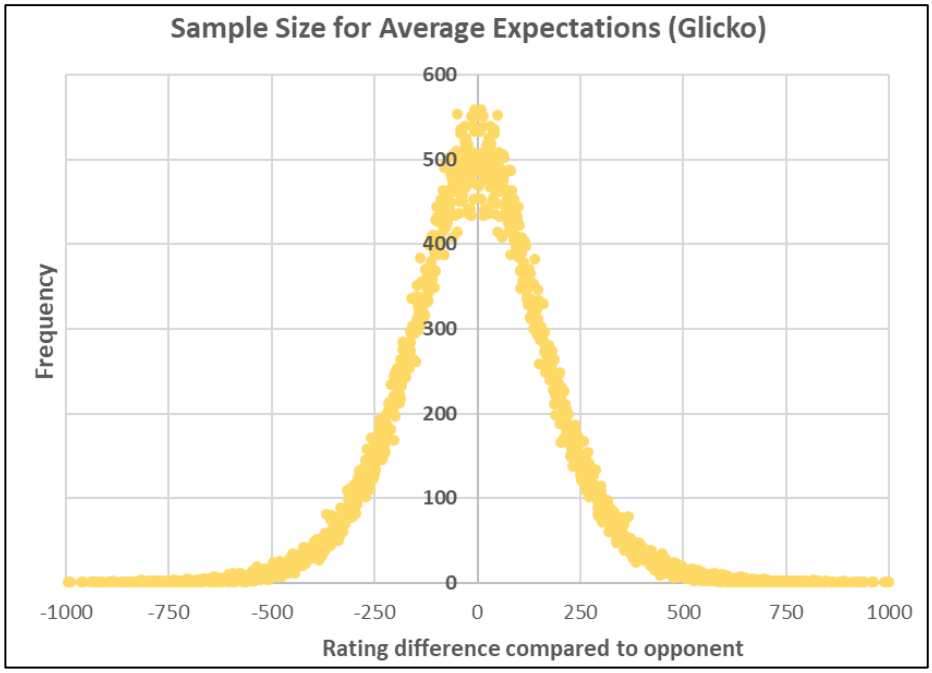


Figure 15 Rating Differences' Frequency Graph for k=230 and parameters listed below

The following parameters were set for the regression solver:

- Both players must be established (at least 50 games)
- Win possibility: 100% (0% to 100%, no cap)
- Player and Opponent Rating range: 950 to 2050
- Player and Opponent RD range: 50 to 150
    - RD = 50 is the minimum level of uncertainty
    - RD = 150 is not exceeded among players who play once every 2 years.
        - Consequently, $0.906 < g < 0.988$ (Table 6)

The solver took the average result ($\frac{total\ actual\ wins}{total\ games}$) at every point of rating difference. Least-square regression is performed, and a new k is determined as best-fit. Each data point is weighted according to its frequency in the 5-year period.

This is sometimes referred to as Error Sum of Squares (SSE), Residual Sum of Squares (SSR) or Total Sum of Squares (SST/TSS) however the models these are applied to are almost always assumed linear. In this situation the authors identified a methodology where they could apply it to the theoretical curve and observed results of each test case, and test for themselves both weighted and unweighted model performances.

This newly determined k was then used to recalculate all game outcomes rating changes from 2006-2019, and again a portion of the most relevant results were analysed and the outcome fed back into the regression solver.

The process would then be repeated again, with minor tweaks, adjusting k to minimise error between expectation and the actual win curve.

Preliminary results from using capped win possibilities of 90% and 95% are not shown because prediction errors improved when these limits were lifted. After removing restriction on win possibility, the solver returns an output k of 250.72 and win predictions are plotted in Figure 16 for comparison against actual win data. Figure 17 shows the detailed prediction errors for each point of rating difference. The average prediction error is 1.12%, a stark improvement over the 2.99% from earlier iterations.

Hence, k = 250 was decided as the final value that determines the win expectation curve's steepness.

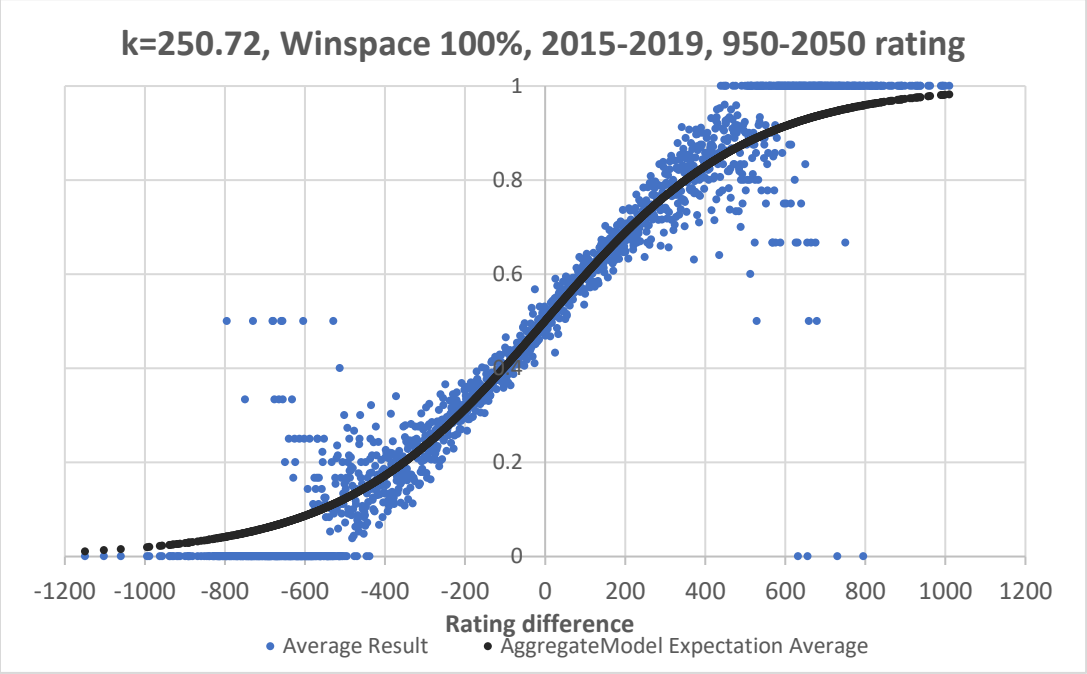


Figure 16 Regression Solver Output imposed on Average Results

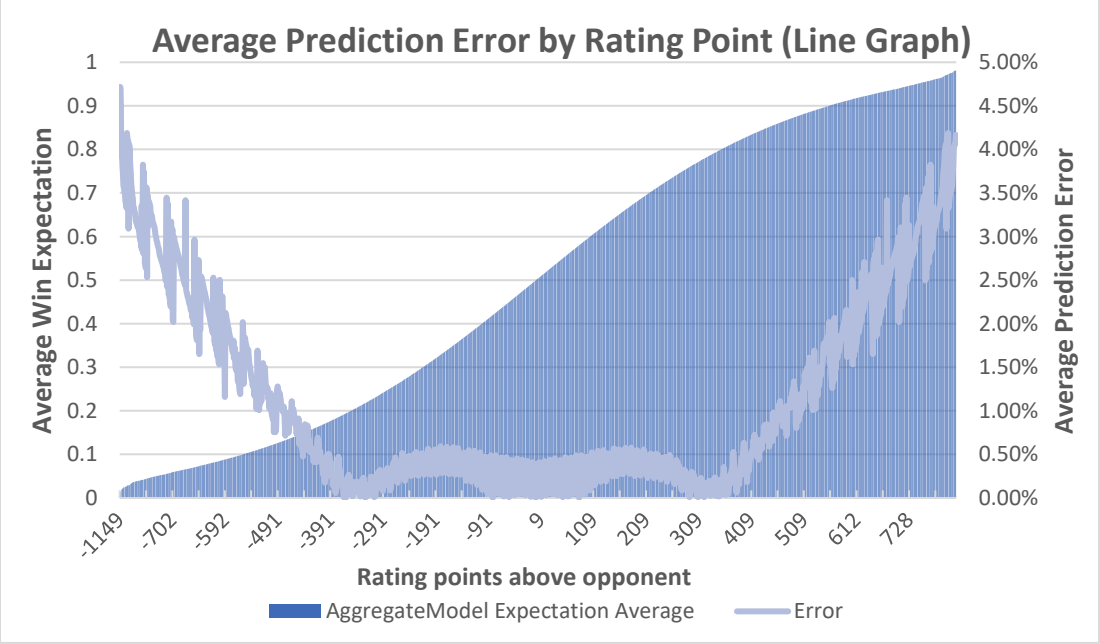


Figure 17 Regression Errors for point-by-point rating difference

The average prediction error is below 1% for rating differences of up to 500 points. With fewer data points and therefore smaller weight at both extremes, prediction errors became larger. To analyse and interpret this prediction error in terms of rating points, the effect of 1% change in win expectation was calculated at multiples of 0.10 in Table 11 below.

The rating increase to produce a 1% expectation difference was the largest at 95% (0.95), requiring an increase of 59 rating points to reach 96% (0.96). That is because the expectation curve has a sigmoid[11] shape that plateaus near 1.00.

As the error is below 100 i.e. twice the minimum RD of 50, it is within acceptable limits of the Glicko rating system. A further result of the system was that it was unlikely for players rated more than 700 points apart ever to meet. Such games account for 142 out of 204,557 analysed game results.

Table 11 Effect of 1% Prediction Error at Various Extents

Expectation	Rating Difference	Error Margin (Rating Points)
0.50	0	±10
0.51	10	
0.60	104	±10
0.61	114	
0.70	217	±12
0.71	229	
0.80	355	±16
0.81	371	
0.90	562	±30
0.91	592	
0.95	753	±60
0.96	813	

[11] Refer to https://www.sciencedirect.com/topics/computer-science/sigmoid-function for a proper introduction to the use of sigmoid functions.

Appendix A - Data Collection Methodology and Challenges

The initial data capture relied on files supplied by WESPA via Chris Lipe and Jason Broersma. Tournament files (.TOU) obtained were tallied with those listed on the WESPA website and matches were found for all tournaments in 2006 and later. 2006 was chosen as the starting point as the World Youth Scrabble Championship began that year, bringing an influx of youth players into the community and rating population.

For visualization purposes, these files were compiled into an SQL database and listed players, individual games, tournaments using related IDs. These IDs were then joined in a relational database to determine which players matched which opponents and in what tournaments. Approximately 16% of all games were pre-2006 and had no matching opponent ID for the given tournament. This was most likely an issue with how past data was compiled and then adjusted in post to allow rating under the current WESPA system. This data quality was deemed acceptable for the purposes of initial analysis, as it represented most of the data accurately.

For rating purposes, the tournament (.TOU) files were sufficient as the code takes all necessary information from .TOU files with no intermediate processing beyond the code.

Once the initial ratings hypothesis of varying k was processed and results analysed for the 2006-2019 test data, the variables were tuned to a closer k value and the methodology was revisited until a close match was obtained between the win expectations and actual wins. In the trials, k was gradually increased from 200 to an eventual 250 determined by a regression solver.

Challenges

1. Opponent IDs and player ID's didn't match 100% of the time.

   Acknowledging that WESPA rating records does contain duplicate profiles for infrequent players, a fuzzy lookup was performed to identify similar looking names and merges were performed for players whose duplicate profiles were identified by or reported to the committee. Treatment for merged profiles is indicated in Appendix C and could be 'Deleted' or 'Absorbed' into existing profiles.

2. Previously rated byes on WESPA record

   Byes were counted towards the games in the past and were sometimes treated as actual games due to errors. These may show up as part of the histogram and graphs. In terms of ratings, these games are ignored through catching opponents named 'Bye [?]' and '[?] Bye', with [?] representing a wildcard with any number of letters. Precautions were taken to ensure one player with 'Bye' in his name would not be removed accidentally.

Appendix B – Data Transition & Handling

1. Changeover Date

In transiting to a new rating system, a starting date has to be decided. The authors' recommendation is that 31 December 2019 is the cutover date for the new rating system. Justification for this had previously been referred to in this paper, with the pandemic's effect on play frequency cited often.

The cutover date implies that WESPA rating history prior to 31 December 2019 is preserved and the new rating data will be referred to beginning on 1 January 2020 to process future rating changes. The new dataset takes all tournament files from 1 January 2006 and has recalculated values from the beginning of time (1 Jan 2006) and an empty[12] ratings list to produce rich rating and RD outputs for every tournament to present.

As a further benefit to this cutover, the following issues identified will also be rectified:

- Byes that have not been caught (i.e. treated as opponents) will be removed. Correspondingly, the player who "defeated" these byes will have their win removed and rating adjusted.
- Duplicate player profiles will be merged to create one consistent profile for the player.
- Unidentifiable player profile such as "Dxxxxx Cxxx" will be removed.

2. Special Case: Rating Changes with 100% unknowns

Some tournaments in the past have had 100% unknown players and have been WESPA-rated. In that case, the ratings code treats all players as 1500 by default. That implies all players will have a win expectation that is 50% of their games played.

Example 7a: Glicko ratings after tournament with all newbies

In this 20-game tournament, all 29 players are new with 1 bye introduced. The newcomer RD of 300 gives rating changes that are about two times larger than normal for the first tournament. Everyone begins at 1500 rating. From there. the points gained per game ranged from 54 points for those with 20 games, to 57 points for those with 19 games. Although there were 29 players, only 15 unique rating outputs are shown in Table 12 as some shared the same number of wins and hence the same rating point.

Example 7b on the next page provides a contrast with the current WESPA ratings.

[12] With an empty rating list, everybody begins at a rating of 1500 by default.

Table 12 Rating Changes for Completely Unknown Players

20 games played, All players are new						
	Expected Wins	**Actual Wins**	**Rating**	**Pre-RD**	**Post-RD**	**Points per game won above 10**
Player 1	10	17	1880	300	136	54.3
Player 2	10	14	1717	300	136	54.3
Player 3	10	13	1663	300	136	54.3
Player 4	10	12.5	1636	300	136	54.4
Player 5	10	12	1609	300	136	54.5
Player 6	10	11	1554	300	136	54.0
19 games played						
	Expected Wins	**Actual Wins**	**Rating**	**Pre-RD**	**Post-RD**	**Points per game won above 9.5**
Player 7	9.5	11	1585	300	139	56.7
Player 8	9.5	10.5	1557	300	139	57.0
Player 9	9.5	10	1528	300	139	56.0
Player 10	9.5	9.5	1500	300	139	0.0
Player 11	9.5	9	1472	300	139	56.0
Player 12	9.5	8	1415	300	139	56.7
Player 13	9.5	7	1359	300	139	56.4
Player 14	9.5	5	1245	300	139	56.7
Player 15	9.5	1	1019	300	139	56.6

Example 7b: Current WESPA ratings after tournament with all newbies

In contrast, the WESPA rating iterates the ratings for a newbie until there is convergence. That means the same tournament result is run many times until the players' ratings produces near-zero rating changes pre- and post-tournament.

Take the three 'Player 7's in Table 13 on the next page as a case study. All three players won 11 games each, but Player 7c got rated 1503 due to more encounters with tougher opponents who became higher-rated **after** this tournament. Player 7a and Player 7b met opponents who became lower-rated **afterwards** and hence their ratings were impacted.

Table 13 Rating changes for tournament with all newbies

	Expected	Actual	Glicko Ratings	WESPA Ratings
Player 1a	10 out of 20 games	17	1880	1959
Player 1b		17	1880	2011
Player 2		14	1717	1717
Player 3		13	1663	1785
Player 4		12.5	1636	1353
Player 5		12	1609	1541
Player 6		11	1554	1581
Player 7a	9.5 out of 19 games	11	1585	1286
Player 7b		11	1585	1289
Player 7c		11	1585	1503
Player 8a		10.5	1557	1075
Player 8b		10.5	1557	1535
Player 9a		10	1528	1205
Player 9b		10	1528	1286
Player 9c		10	1528	1503
Player 10		9.5	1500	1066
Player 11 (3 persons)		9	1472	920
		9	1472	1298
		9	1472	1151
Player 12 (showing 3 out of 6)		8	1415	1042
		8	1415	935
		8	1415	887
Player 13		7	1359	874
Player 14		5	1245	540
Player 15		1	1019	300

Although all these players began the tournament equally as newcomers, their post-tournament rating has returned to influence pre-tournament ratings. This is an undesirable outcome. That has led to some staggering rating differences among players who have the same number of wins, sometimes differing by 300 to 400 rating points (highlighted in red) despite achieving the same results.

The new system addresses this unfairness by setting a default of 1500 for this all-newcomer scenario. However, the ratings officer may decide on an initial starting point different from 1500 if (s)he believes it would provide a more realistic estimate of skill levels. This could be adjusted in steps of 50, based on criteria such as:

- Presence of age restrictions for category (U-12, U-15, U-18 etc.)
- Newness of country
- Other player ratings from same country

Appendix C – Visualizations of Player Performance

In Section 2.2 Paragraph 1, actual wins and expected wins were presented in bins of 25 and 50 rating points through Figure 5**Error! Reference source not found.**. The graph below s hows the result when there is no grouping done. For rating differences > 500 points, the sample for each point (500, 501, 502, etc.) of difference was sparse, hence an outlier performance would sway the win rates to values such as 0.5 (1 out of 2) or 0.333/0.667 (1 in 3, 2 in 3).

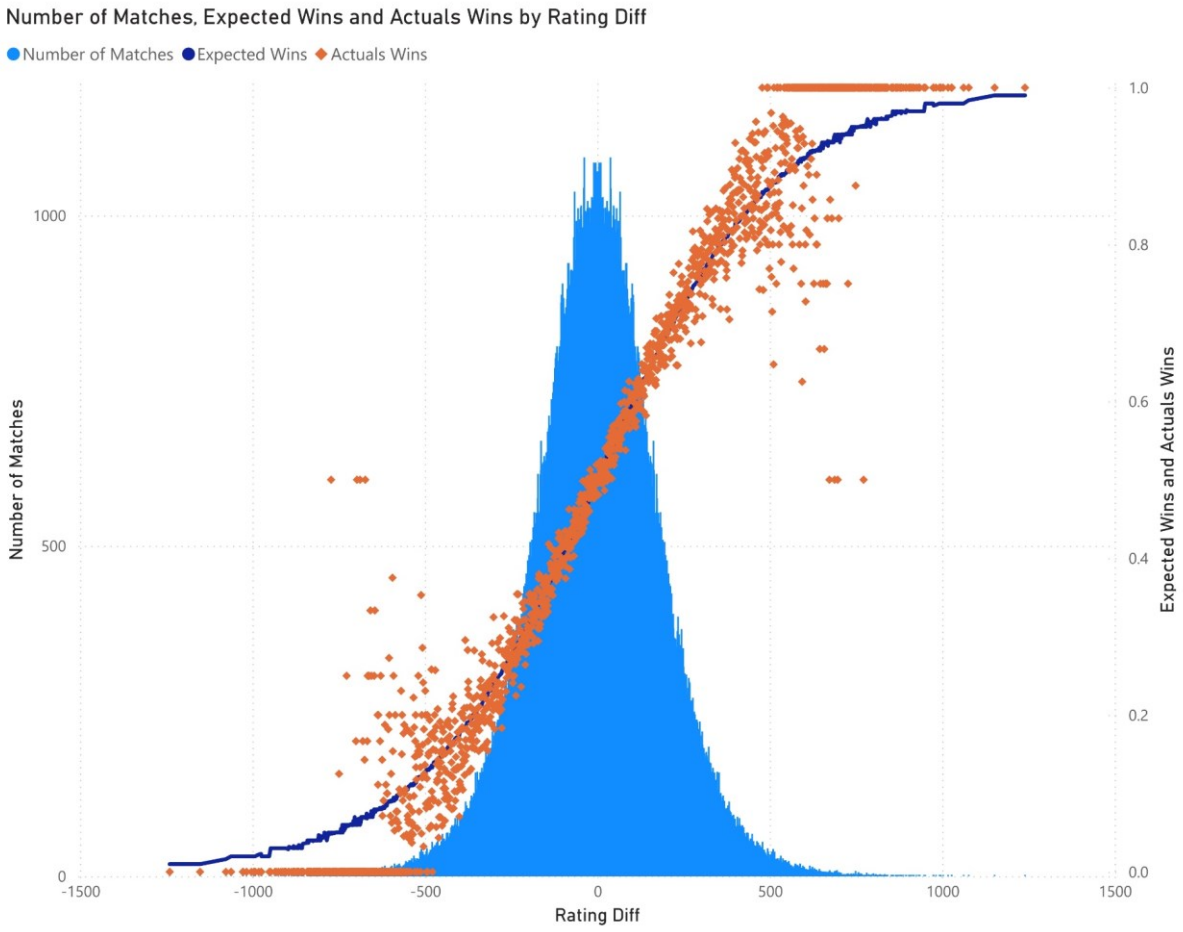


Figure 18 Expected Wins vs Actual Wins by Individual Point (Glicko, k=250)