Ratings System Review

Interim Report

Prepared by: Steven Brown

May 2016 (minor edits June 2016)

RATINGS SYSTEMS REVIEW

Background

The management meeting at the 2014 Nationals asked Glenda Foster to convene a panel to review the ratings system. Glenda advised the management meeting at the 2015 Nationals that she had nothing to report, and felt that I would be in a better position to review the system. So, over the last year I have been reviewing the system and posting material to the #ratings-review channel at nzscrabble.slack.com, seeking feedback from others interested in the review. I'd like to particularly thank Chris Handley for his input.

Recommended Goals of Review

In my opinion, the following general goals should be preferred for the review.

- Continuity with the current system, unless the current system is irretrievably broken.
- Changes that address issues, i.e. that improve the current system, but don't introduce new problems.
- Aim for as simple a system as possible, keeping in mind the axiom that "perfect is the enemy of good".
- Aim for consistency with aspects of other systems, where possible.

Approach to Review

My understanding is that it was intended that the first question the review should address was whether there were particular problems with the current system which could or should be addressed. Then (if it was found that there were such problems with the current system) were they best addressed by modifying the current system, or by starting fresh with an entirely different system. I started by collecting various criticisms of aspects of the current system, and analysing these in the light of other rating systems, and by analysing the data in the ratings database for the current system.

Rating System Theory and History

Our ratings system is an ELO-style ratings system (as are the world, North American, and Australian systems). The UK uses a somewhat different system. The main features of an ELO-style system are that the probability of winning each game is estimated according to the difference in rating between each two opponents by looking it up on a curve, and that the difference between this expected win probability and the actual wins achieved (multiplied by a "k-factor") is used to adjust the player's rating. The k-factor represents the importance the current set of results is given compared to all previous results, as represented by the original rating.

ELO-style ratings systems started off being used for chess (courtesy of its creator, Mr Arpad Elo). The original system used the cumulative normal distribution (the inverse of the probit function) as its probability curve. Some time later, the chess system started using the logistic function (the inverse of the logit function), scaled to closely approximate the previous curve, instead. My understanding is that the cumulative normal distribution better represents the probability of unconstrained outcomes, while the logistic curve is better for constrained outcomes. This is the difference between, for example, flipping one coin to decide which player wins a game (constrained to one win each game), and flipping one coin for each player (unconstrained - each game could have two winners, two losers, or one of each).

Our current system uses the cumulative normal distribution (scaled using 200 times the square root of 2). The old Australian system, the old 'world rankings' system, (and I suspect the old NSA system) used a logistic curve scaled (using 172) to closely approximate our current curve. I suspect both these curves came straight from chess at different times. Over time, various sets of real-world Scrabble results have been analysed, and ratings curves have been changed as a result. In 2006, Australia changed to a straight-line curve with a slope of 1 in 12. In 2009, the NSA kept the logistic curve but changed the scaling from 172 to approximately 313. In 2012, WESPA adopted a new system also using a logistic curve scaled using 313 (having also considered as a final contender a straight-line curve with a slope of 1 in 15).

Analysis of ratings database

We have no comprehensive record of the results of individual games (apart from the scoring files from occasional tournaments), so we can’t aggregate all the games between players with a particular ratings difference and work out the average win percentage at that level of detail.

But we do, in the ratings database, have records for every tournament since the current system was introduced that let us work out what the expectancy (i.e. expected win percentage over the tournament) of each player must have been, and compare this to the win percentage achieved by that player. This means that individual rating differences are aggregated according to the application of the current ratings curve, but hopefully the large volume of data (representing about 10,000 games per year, on average, over the last 17 years) means that the results are still meaningful.

We can work back from the expectancy to a ratings difference. This is not the average ratings difference with the opponents played, unless all of the opponents of the player had the same rating, but it’s the best we have. We can then work out the average actual win percentage for a set of ratings bands (I took each percentage of expectancy under the current system as a ratings difference band). We can then work out for each of those bands what the win expectancy would be under a variety of other ratings curves. The results of this analysis were as follows (a bigger number means the curve in question was cumulatively ‘further away’ from the actual win percentages.

Ratings curve	Cumulative difference
Current NZ system (200 * sqrt(2) scaling, flat cutoffs at 0.5, 0.95)	4.07
Old Australian and WESPA curve (logistic with 172 slope)	4.03
Current system with regression to mean	1.22
Current Australian system (straight line, 1/12 slope, with flat cutoffs)	0.89
Considered for new WESPA system (straight line, 1/15 slope, with flat cutoffs)	0.66
Current WESPA and NASPA curve (logistic with 313 slope)	0.46
Current system (with scaling changed to 365 * sqrt(2)	0.43

It is possible to get the logistic curve down to about 0.44 (with a slope around 320), but this kind of fractional difference is not enough to outweigh other factors.

One factor to consider is the ease of implementation of the system in software. The functions used to calculate the CND curve in the current NZ system are only easily available in a Microsoft Excel spreadsheet; in most programming languages, they are only available in separate statistics packages and the appropriate parameters to use are not always obvious. The functions used to calculate the logistic curve (exponents of e and base-e logarithms) are part of the core Maths package in most programming languages, and are available on any calculator in scientific mode. A straight-line curve would be easiest to calculate, but, as can be seen above, would not be as accurate.

Issues with Current System

- **The current ratings curve is too steep (as evidenced by often seeing majority of players in top of grade losing points and majority of players in bottom of grade gaining points (this is caused by the 'luck factor', and by using the chess system, which doesn't account for it)).**

This seems borne out by the above analysis - adopting the NASPA/WESPA curve would lead to lower expectancies at the top of each grade, and higher at the bottom. If the historic pattern of win percentages at a given ratings differential continues as ratings differentials change under the new system, this should lead to a more even distribution of gains or losses of ratings points for different seedings in grades.

The 'luck factor' is just part of the overall variability in results that this kind of system is designed to cope with. If this variability wasn't present in chess, chess-players wouldn't need to play a long series of games to decide a championship. The 'luck of the draw' present in Scrabble certainly increases this variability, but it's a difference of degree not of kind. This is why we need to scale the slope of the curve differently, but don't need to use a completely different curve.

- **The current ratings system is too volatile (as evidenced by players, especially in lower grades, and provisionally-rated players, gaining or losing hundreds of points at a time).**

Our current k-factor varies from 60 points per game different to expectancy at a rating of zero to 0 points at a rating of 3000. The NASPA system varies from 30 points (for less than 1800 rating and less than 50 games played) to 10 points (for 2000+ rating and 50 or more games played). We could get close to this range by simply halving our current k-factor.

- **One day tournaments too short to be meaningful for ratings - contribute to volatility.**

This could be addressed by way of the k-factor. Rather than divide the current k-factor by 2, we could divide it by 20, then multiply by the number of games being rated. Doing this would make the k-factor at a one-day tournament about 1/3 of current k-factors, at a 2-day tournament, about 3/4, and slightly more than at present at the Masters. Back-to-back one-day tournaments could be required to be rated together.

- **It is possible to win your grade and still lose ratings points (or to 'lose' your grade but gain ratings points).**

If you consider that in a one-day round robin, it is possible to 'win' your grade if all the players win half of their games (3 1/2 wins), and you have the best spread; or if you win all 7 games. In the first case, your 'win' relies almost entirely on how the other players in the grade went against each other, which you shouldn't get credit for in your rating. The prize table is there to reward you for winning your grade - leave the ratings system to make the best guess it can as to how many games you might win next time.

It should be true in the current system that you won't win your grade and go down in the rankings past anyone else in the grade. You might go down past someone who hasn't played in order to 'protect' their rating, but this is a problem that is best addressed by other means.

- **The more players that achieve part of their expectancy, the harder it is for all of them to achieve the balance of their expectancy.**

Situations like this might make problems with the steepness of the ratings curve more obvious, but the phenomenon itself is true of any rating system - it's always going to be less likely that any two people get their expectancy than that any one person does. And even less likely that three all get their expectancy.

- **Players can spend too long with a provisional rating - provisional rating is very volatile right through to first established rating.**

The current WESPA system has 30 games (compared to our 35) as number of games someone has a provisional rating for. These 30 games are collected, and rated as a group, whereas, in our system, the only affect an old provisional rating has on the new rating is the grade it put you in. The NASPA system gives someone an initial rating at their first tournament, and they are then rated the same as anyone else (with a higher k-factor for the first 50 games). As our current provisional calculation is the inverse of our current ratings calculation, adopting a shallower curve would, while probably making established ratings less volatile, make provisional ratings even more volatile.

- **Treatment of players with historic (and possibly stale) rating different to players new to NZ rating system.**

This needs rationalising. I suggest calculating an initial rating (using something like our provisional formula) for the first tournament, as in the NASPA system, then classing the rating as provisional for the first 30 tournament games (as in the WESPA system). Provisional ratings would then be rated the same as other ratings, but tournament organisers would have discretion as to which grade to place any players with new, provisional, or historic ratings. The details would need testing during implementation of the new system.

- **Rapidly improving players take points from those they play (and not from those they don't).**

This is not the case with our current provisional rating system (unless there is more than one provisionally-rated player in the same grade). This issue would need to be considered as part of deciding on the exact implementation with regard to new, provisional, and historic players.

- **Ratings are not comparable over time (mostly due to ratings deflation).**

As ratings are derived entirely from the relative results of players, they can't be expected to be consistent over time. What should be relatively consistent over time is the ranking of players as ratings as a whole go up and down. This relies on as many players playing as many other players as possible, as often as possible. The nature of our k-factor helps a little to keep ratings comparable, but this could be augmented.

- **Players avoiding tournaments to protect their rating distorts whole rating system.**

One way to address this problem would be to add a small number of points to the new ratings of players after a tournament (say, one point per game rated). The points would then be up for grabs the next time the player played. The nature of our k-factor should keep a lid on the maximum rating, and help make ratings more comparable across time in future.

Recommended Course of Action

Adopt NASPA/WESPA curve (close enough is good enough, ease of development).

Modify K factor from (3000-currentRating)/50 to (3000-currentRating)/1000*ratedGames

Allow back-to-back one-day tournaments to be rated as if they are one two-day tournament.

Derive initial rating from first tournament only (in line with NASPA).

Reduce provisional games to 30 (in line with WESPA).

Allow all New, Provisional, Historic players to be re-graded (with agreement).

Rate all New, Provisional, Historic players on the same basis as established players.

Add ‘participation points’ to each new rating (one point per game rated).

Start new system from first tournament next year.

Feedback Sought on Review

- Agreement or not on adopting NSA/NASPA/WESPA curve.
- Agreement or not with modifications to K factor.
- Agreement or not with general approach to new/provisional/historic ratings.
- Agreement or not with participation points.
- Start in Jan 2017 - should previous years (up to 5) be re-rated using new system?