Listening. Learning. Leading.®

Research Memorandum
ETS RM–15-03

Ranking Systems Used in Gaming Assessments and/or Competitive Games

Ourania Rotou

Xiaoyu Qian

Matthias von Davier

July 2015

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Ranking Systems Used in Gaming Assessments and/or Competitive Games

Ourania Rotou, Xiaoyu Qian, and Matthias von Davier

Educational Testing Service, Princeton, New Jersey

July 2015

Action Editor: James Carlson

Reviewers: Andreas Oranje and Alina von Davier

Abstract

With the advances of technology, educational assessment can be further developed and provide more comprehensive information about examinees. Creating a testing environment where examinees can be engaged in a more collaborative, interactive, competitive, and enjoyable way not only can increase examinees' motivation but also can provide measures of skills and cognitive abilities that traditional assessments do not. The feature of mimicking a realistic setting using a virtual environment makes gaming assessment an attractive tool for collecting data about the performance of examinees and actually selecting examinees for advanced social roles and occupations that require multiple abilities and skills. Gaming assessment and other competitive games apply ranking systems to rank candidates. In education, the application of ranking systems is not used, even though the ranking of examinees is a common practice. For example, stakeholders rank universities, administrators use test scores to admit students to colleges, and teachers rank students to place them in advanced tracks. In this paper, we review several ranking systems that have potential application in educational assessment. These ranking systems include models that rank two-player systems such as Ingo, Elo, Glicko, and Edo and multiplayer systems such as TrueSkill.

Key words: gaming assessment, pairwise comparison, Bradley-Terry model, chess ranking, Elo system, Trueskill

Corresponding author: O. Rotou, E-mail: ORotou@ets.org

Educational assessment is a process of collecting, analyzing, and determining in measureable terms examinees' status on knowledge, skills, and attitudes. Traditionally, the outcomes of educational assessment are test scores that represent the performance of a test taker on a test that measures constructs of interest. Test scores can be used in different ways. For example, some test scores are used to rank test takers compared to the population of test takers (norm-referenced tests) or to determine whether a test taker has learned the content that is relevant to the construct of interest (criterion-referenced test.).

With the advances of technology, educational assessment can be further developed and provide more comprehensive information about test takers. Creating a testing environment in which test takers can be engaged in a more collaborative, interactive, competitive, and enjoyable way not only can increase their motivation but also can provide measures of skills and cognitive abilities that traditional assessments do not. For example, gaming assessment can model a more complex virtual setting in a lab or in a research field, such as a deep ocean, an archaeological site, a space station, and so on. The feature of mimicking a realistic setting using a virtual environment makes gaming assessment an attractive tool for collecting data about the performance of test takers and actually selecting them for advanced social roles and occupations that require multiple abilities and skills. These occupations can be found in fields such as education, science, technology, engineering, military, aviation, or medical research.

Norm-referenced tests have been used widely in the educational field. For example, administrators use test scores to admit students into colleges. The selection of test takers is a direct process of ranking them based on test scores derived from traditional educational assessments. Although ranking models are not widely used in the educational field, these models have wide applications, including ranking of Web sites, players of tennis and chess, items for purchase, or stimuli in psychophysics (Langville & Meyer, 2012).

Many ranking systems originated with modern chess ranking systems as early as the 1930s (Chess rating system, n.d.), and more recently are applied widely in gaming assessment, for example, the Microsoft Xbox video game system. Competitive games such as chess use skill rating systems for several practical purposes: (a) to qualify candidates for elite tournaments, (b) to pair candidates of similar abilities for tournaments, and (c) to monitor candidates' progress (Glickman, 1999).

In general, rating systems are designed to provide information about players' skill development by combining information from a new game outcome with players' skills as demonstrated from previous games. These systems aim to provide information about a player's strength at any time. Some systems update players' strength estimates after each game, whereas others update information after a series of games. Some of these systems are developed to rank two-player games, and in more recent years, ranking systems have been further developed to rank players in multiplayer games.

The purpose of this document is to provide information about some of the most well-known existing ranking systems. This paper is organized in two parts. The first part summarizes, compares, and contrasts some of the best known ranking systems for two-player games, and the second focuses on the TrueSkill system (Herbrich, Minka, & Graepel, 2006), which is intended for multiplayer games.

Skill Ranking Systems for Two-Player Games

In this section, we describe and identify similarities and differences among the most well-known ranking systems. They are the Ingo, Elo, Glicko, and Edo systems. Further details for each system are available in the appendices.

Ingo System

One of the first ranking systems to produce numerical ratings, the Ingo system was developed by Anton Hoesslinger in 1948 and used by the German Chess Federation. Over the course of the following decade, many versions of this system were developed and used in different national chess tournaments. The Ingo system was used for paired comparisons. Unlike other systems, the Ingo system associates better performance with lower scores.

The Ingo system is considered a simple one, with little basis in statistical ratings. A player's ranking is based on the performance of the average player. In particular, the average rating of the players in a competition is calculated. Also, the player's score in percentage points is calculated. If a player's percentage score is average (50%) then the player's rating score is the average rating score; if the player's percentage score is above 50%, then the player receives the average score plus 10 points for each percentage point above 50%. Similarly, if the player's percentage score is below 50%, then the player receives the average score minus 10 points for each percentage point below 50%. For example, if the average rating score in the competition is 1,500 and the percentage score of a player is 23%, then this score is 27 percentage points below average, so the new rating score of the player is 1,500 – (10 * 27) = 1,230.

Elo System

The Elo system was developed by Arpad Elo in 1959 and adopted by the World Chess Federation in 1970 (Elo, 1986; Elo rating system, n.d.). It is probably the most widely used system in competitive games such as chess. Like the Ingo system, the Elo system is a ranking system for two-player games. However, the Elo system is based on a model with considerably more statistical foundation. The Elo system assigns a number between 0 and 3,000 that changes over time based on the outcomes of tournament games. Unlike the Ingo system, in the Elo system, a higher score indicates better performance. Thus, a player with a higher rating is expected to win more often than a player with a lower rating. Based on the game outcomes, the player's rating may be increased or decreased.

The main assumption of the Elo system is that each player is associated with a current strength, and this strength is estimated by a rating. The Elo system associates game results to latent variables that represent the ability of each player. The Elo system uses the Thurstone-Mosteller model to estimate the probability of individual game outcomes based on the assumption that the player's chess performance in each game is a random variable that is normally distributed. It is assumed that true ability of each player is the mean of that player's performances. Performance is measured by wins, losses, and draws.

The assumption that a player's performance is normally distributed raises some concerns. Some statistical tests have indicated that this assumption does not accurately represent the actual results, especially for weaker players, who have greater chances to win than Elo predicts. For this reason, some chess sites use a logistic distribution. The logistic distribution version of the system goes back to Zermelo (1929), who developed a model for paired comparisons that later became known as the Bradley-Terry model (Bradley & Terry, 1952; Luce, 1959). The Bradley-Terry model is an approach to ranking *n* individuals by comparing two at a time.

One of the greatest assets of the Elo system in terms of usability is its linear approximation. The linearization of this model makes it attractive to users due to its simplicity. If players win more games than expected, their ratings will increase. Similarly, if players lose more games than expected, their ratings will decrease. However, the adjustment is assumed to be linearly related to the number of wins/losses by which the players differ from their expected number of wins/losses. Furthermore, players' performance ratings are a function of the opponent rating and a linear adjustment to the amount by which they overperform or underperform their expected values. All things being equal, when players' actual scores are less than the expected values, their ratings are adjusted downward. On the other hand, if their actual scores are higher than their expected scores, the ratings are adjusted upward. The rating update for each player can be performed after each game or after a defined rating period.

Although the linear nature of this model makes it simple, advances in technology have made it obsolete. One of the limitations of the simplicity of the Elo model is that more efficient estimation models are becoming more attractive. Another limitation of the Elo model is that it uses a player's most recent rating as the current one, even if the player has not competed for a long time.

Glicko System

The Glicko system (Glickman, n.d.) was developed by Glickman in 1999. Like the Ingo and Elo systems, the Glicko system is designed for two-player games. This model is an extension of the Elo system and was developed in an attempt to address and improve the parameter estimates by incorporating a variability factor. The Glicko system computes the rating in a similar fashion to the Elo system, but it also incorporates the reliability of a player's rating. The reliability of a rating is called the rating deviation (*RD*), which is a standard deviation that measures the uncertainty of the rating. For example, a player who did not play for a long time and had just one game may have a high *RD*. A player who competes very often may have a low *RD*. The rationale is that the system can gather more information about the skill of the player who competes more often, and therefore the rating is more precise than that of a player who competes less often. Because the Glicko system provides both a rating and an *RD*, it may be more informative to describe players' skills as a confidence interval. For example, a 95% confident interval is calculated as Rating ± 2 * *RD* (Glickman, n.d.).

According to Weng and Lin (2011), the Glicko system was the first to use the Bayesian ranking system. It is assumed that the skill of the players follows a Gaussian distribution. The Glicko system basically applies the Zermelo model (Zermelo, 1929), better known as the Bradley-Terry model (Bradley & Terry, 1952; Luce, 1959). As mentioned earlier, the Bradley-Terry model is an approach to rank *n* individuals by comparing two at a time. The Glicko system updates the skill of the players after each rating period. For better estimates, the number of games in each rating period is between five and 10 games for each player (Weng & Lin, 2011).

A drawback of the original Glicko system (Glicko-1) is that it may not capture the true change in skills for players who compete frequently. This is because the *RD* is small for players who compete very often. As a result, the rating for these players may not change accurately (Glickman, n.d.).

In addition to the Glicko-1 system, Glickman developed the Glicko-2 system. The Glicko-2 adds a rating volatility to the rating and *RD*. The rating volatility index is the degree of expected fluctuation in a player's rating. The volatility measure is low when a player has consistent results, and it is high when a player has inconsistent performance. As with the Glicko-1 system, results for the Glicko-2 system are updated after a rating period. Like the Glicko-1 system, Glicko-2 performs best when rating periods consist of five to 10 games for each player. It should be noted that the ratings outcomes based on the Glicko-2 system are very similar to the ones from the Glicko-1 system, because the outcomes do not incorporate any evidence of the volatility index (Glickman, 2013).

Edo System

The Edo rating system has been developed and maintained by Rod Edwards since 2004 (Edwards, n.d.). Similar to the systems discussed above, Edo is a rating system for paired comparisons. In addition, like the Glicko system, Edo is also based on the Bradley-Terry model (Zermelo, 1929). Its mean rating is adjusted to roughly 1,500 with a standard deviation around 300.

What makes the Edo system distinctive is that during the rating/estimation, the system treats the same player at two different years as two different players. The rating of players who participated in matches in two different years is then computed as a weighted rating between the 2 years, as if the players had played against themselves in those years. The weight is set up around 50%. A weight higher or lower than 50% can compensate for inflation or deflation of the rating from time to time (e.g., due to a player's skill increase). Also, according to Edwards (n.d.), more self-matches of the same player result in a more stable rating of the player, whereas fewer such games mean that the player's rating is more the result of current performance.

Because at the end of 20th century more local tournaments with players at the lower end of the rating skill were included compared to earlier times, there is a tendency during modeling for estimation to be pulled down when more local tournaments are recorded. The second distinctive factor of the Edo system is that an adjustment is made to account for this situation: Players with ratings higher than 1,500 are marked down, while players with ratings lower than 1,500 are elevated. After this adjustment, the maintained result is similar to that of the Elo system.

In addition, Edwards (n.d.) also claimed that the Edo system has advantages in measuring uncertainties when compared to the Glicko system. For example, when a small group of players has played against one another but not often against players outside of the group, the Edo system has “some links” (see Edwards, n.d., “Measuring uncertainty - rating deviations” on page http://www.edochess.ca/Edo.explanation.html) to the main group under this situation. However, it is unclear how these links are maintained and estimated. Further, although this model considers information for the same player at different times and provides variance of the player’s skill, it does not provide posterior distributions, is not a full Bayesian model, and does not model draws (Dangauthier, Herbrich, Minka, & Graepel, 2007).

Skill Ranking Systems for Two-Player and/or Multiplayer Games

The TrueSkill Ranking System

In the second part of this document we give an overview of the TrueSkill rating system. The TrueSkill model was developed by Microsoft Research (Herbrich et al., 2006) and may be viewed as a generalization of the Elo system to multiplayer games. The TrueSkill ranking system is used for Microsoft’s Xbox online games, and in general, it is used to rank players for video games with more than two players and/or teams per match in competitive games. The simplest scenario for TrueSkill is the same as the one described in the Elo and the Glicko systems for two players competing against each other. However, the TrueSkill model was reported to provide more accurate estimates in predicting game outcomes and in matching players compared to the Elo system (Herbrich et al., 2006).

The TrueSkill system uses Bayesian approximation estimation (Kschischang, Frey, & Loeliger, 2001; Minka, 2001), which allows for instantaneous ranking updates of players and/or teams after each game. In a game, each player is assumed to have a prior skill with a mean and a standard deviation, and a Gaussian distribution is assumed. In Xbox Live, a prior skill with a mean of 25 and a variance of $(25/3)^2$ is used for the initial run. The performance of players in a game has a mean around their estimated skill with a standard deviation. The performance of a team is the sum of each member’s performance. Each team’s performance is then compared to decide team ranking. Draws (players with equal ranks of performance) are allowed in the TrueSkill ranking system.

If the difference between two teams in terms of their performance is less than a draw margin, these two teams are ranked at the same level. The draw margin can be narrow or wide depending on the needs of the estimation. A narrow margin should be used when individuals’/teams’ skills are relatively close and it is important that fewer ties are observed in the ranking. On the other hand, a wide margin should be used when ranking is more entertaining and low stakes. Posterior estimation of each player’s skill is then used as a prior for ranking estimation of the player’s next game. The estimation algorithm of TrueSkill uses approximate message passing—a Bayesian approximation method (Kschischang et al., 2001; Minka, 2001). It is reported that convergence is fast; thus, instantaneous ranking is possible (Weng & Lin, 2011).

The initial TrueSkill rating system ranks game players at a certain time point (*t*) by updating their earlier rankings (*t*-1) as the prior and always estimates players’ rankings forward through time. Dangauthier et al. (2007) extended TrueSkill to estimate players’ skills not only forward through time but also backward. They called this extension TrueSkill Through Time (TTT) or TTT-D when the estimation of an additional draw margin parameter discussed earlier is included. Under TTT, for example, if Player A beats Player B, and then later, Player B beats a strong Player C, TTT and TTT-D are able to adjust Player A’s ranking by going backward in the estimation. However, the original TrueSkill rating system is not able to make the backward adjustment for Player A in this case.

Dangauthier et al. (2007) claimed that TTT and TTT-D are more accurate ranking systems than the original model. However, a longer estimation time is required and inevitable because there are more steps in the algorithm when estimation goes forward or backward in order to consider ranking of players who were rated previously, and adjustment is needed when new players are lined up to be ranked. In an experimental run, the TTT rating system was used to rank chess players over a 150-year time span. The estimation of TTT took around 10 minutes, and that of TTT-D took around 20 minutes on a Pentium 4 machine. Although the authors of TTT and TTT-D claimed that their algorithm is more accurate, they applied these two rating systems only to the two-game player scenario of chess rating data. It is unclear whether these models would apply to the Xbox multiteam, multiplayers rating scenario. The authors have shared the data and code of the ranking system (Herbrich & Graepel, 2008).

An important feature of the TrueSkill ranking system is player matchmaking (Graepel & Herbrich, 2006). In order for players to have a competitive and enjoyable gaming experience, skills of competitors have to be close. TrueSkill is able to match online players based on their estimated skills. There are two scenarios: games of individuals and games of teams. In a multiplayer (nonteam) game, a simple criterion used for matchmaking is to ensure that the players' highest and lowest ratings in a game do not go above a predetermined rating difference. In a multiteam game, a team member's ranking is estimated with all the other players to get a pairwise rating. For each player, relative pair standings are then averaged as the player's ranking. The criterion for multiteam game matchmaking is to have about the same number of players on each team and also for all team players across teams to have similar skill levels.

In addition to the original TrueSkill model, several TrueSkill variant models have been used for online data: multilabel classification (Zhang, Graepel, & Herbrich, 2010) and Web commercial click rate prediction for Microsoft's Bing search engine (Graepel, Candela, Borchert, & Herbrich, 2010).

Wide use of TrueSkill/Bayesian approximation algorithms suggests that model building similar to the TrueSkill rating system is flexible and that the Bayesian approximation algorithm is useful in real-world applications. Using a Bayesian approximation algorithm presents several advantages. First, because video game ranking is online and instantaneous, as compared to traditional frequentist batch data estimation, Bayesian approximation fits into the scenario very well. Players' skills can be updated within a short period of time after each game based on their prior skills. Second, Bayesian approximation is a compromise between estimation resources (time and cost) and accuracy. While Bayesian approximation may be less accurate than fully Bayesian models, the nature of online ranking does not allow for the long estimation time and high computational cost required by the latter.

Scoring methods are a vital feature of assessment (e.g., educational assessment, competitive games, gaming assessment, and so on). This paper provides information about ranking systems in gaming assessment and competitive games. These ranking systems can be further considered and examined in educational assessments, especially in the context of norm-reference tests in virtual environments. Use of efficient and effective scoring methods in digital environments that provide immediate feedback could enhance learner’s motivation, monitor learning, and provide more comprehensive information about cognitive and noncognitive skills of test takers.

References

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39,* 324–345.

Chess rating system. (n.d.). In *Wikipedia*. Retrieved April 27, 2015 from http://en.wikipedia.org/wiki/Chess_rating_system

Dangauthier, P., Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill through time: Revisiting the history of chess. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems, Vol. 20* (pp. 337–344). Cambridge, MA: MIT Press.

Edwards, R. (n.d.). *Edo historical chess ratings.* Retrieved from http://www.edochess.ca/index.html

Elo, A. E. (1986). *The rating of chessplayers, past and present* (2nd ed.). New York, NY: Arco.

Elo rating system. (n.d.) In *Wikipedia*. Retrieved April 27, 2015 from http://en.wikipedia.org/wiki/Elo_rating_system

Glickman, M. E. (n.d.). *The Glicko system*. Retrieved from http://www.glicko.net/glicko/glicko.pdf

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics, 48,* 377–394.

Glickman, M. E. (2013). *Example of the Glicko-2 system.* Retrieved from http://www.glicko.net/glicko/glicko2.pdf

Graepel, T., Candela, J. Q., Borchert, T., & Herbrich, R. (2010). Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning 2010* (pp. 13–20). Retrieved from http://research.microsoft.com/pubs/122779/AdPredictor%20ICML%202010%20-%20final.pdf

Graepel, T., & Herbrich, R. (2006). Ranking and matchmaking: Grouping online players for competitive gaming. *Game Developer Magazine, 13*(9), 25–34.

Herbrich, R., & Graepel, T. (2008, April 5). TrueSkill through time [Web blog post]. Retrieved from the Applied Games Group Blog, http://blogs.technet.com/b/apg/archive/2008/04/05/trueskill-through-time.aspx

Herbrich, R., Minka, T., & Graepel, T. (2006). TrueSkill^{T}: A Bayesian skill rating system. *Advances in Neural Information Processing Systems, Vol. 19* (pp. 569–576). Cambridge, MA: MIT Press.

Kschischang F. R., Frey, B. J., & Loeliger H. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory, 47,* 498–519.

Langville, A. N., & Meyer, C. D. (2012). *Who is #1? The science of rating and ranking.* Princeton, NJ: Princeton University Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

Minka, T. P. (2001). A family of algorithms for approximate Bayesian inference (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, Massachusetts.

Weng, R. C., & Lin, C. (2011). A Bayesian approximation method for online ranking. *Journal of Machine Learning Research, 12,* 267–300. Retrieved from http://jmlr.csail.mit.edu/papers/volume12/weng11a/weng11a.pdf

Zermelo, E. (1929). Die berechnung der Turnier-Ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung [The calculation of tournament results as a maximum likelihood problem]. *Mathematische Zeitschrift, 29,* 436–460.

Zhang, X., Graepel, T., & Herbrich, R. (2010). Bayesian online learning for multi-label and multi-variate performance measures. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics: Vol. 9* (pp. 956–963). Retrieved from http://research.microsoft.com/pubs/122783/zhang10b.pdf

Appendix A
Ingo System Overview

Historical Information

- Developed in 1948 by Anton Hoesslinger
- Adopted by the German Chess Federation
- First chess rating system
- In the decade after its development, several versions of this system were developed.

Comparisons

- Two-player comparisons

Statistical Information

- Has little basis in statistical theory
- Calculates player’s ranking based on the performance of the average player
- Lower scores indicate higher performance.

Advantages

- Simple model to use

Limitations

- Ratings were consistent with subjective ranking of chess players.
- A player could lose in every game and still gain rating points.

Appendix B
Elo System Overview

Historical Information

- Developed in 1950 by Arpad Elo
- Adopted by the World Chess Federation in 1970
- Developed as an improved rating system over the Ingo system
- Most widely used system in competitive games

Comparisons

- Two-player comparisons

Statistical Information

- It is based on a model with considerably more statistical foundation compared to the Ingo system.
- The performance rating of a player is a function of the opponent rating and a linear adjustment to the amount by which a player overperformed or underperformed that player's expected value. All things being equal, when a player's actual score is less than that player's expected value, the rating is adjusted downward. On the other hand, if the actual score is higher than that player's expected score, the rating is adjusted upward.
- Higher scores indicate better performance. For example, when two players compete, the system predicts that the player with the higher rating is expected to win more often than the player with the lower rating.
- It uses two different distributions. It assumes that players' performance distribution follows either a normal or a logistic distribution.

Advantages

- Applies Thurstone-Mosteller model
- The range of the rating scores is between 0 and 3,000.

Limitations

- The rating update for each player can be performed after each game or rating period.
- The linearization approximation of this model makes it attractive to users.
- With the use of technology, models with more efficient estimations are becoming more attractive.
- Uses player's most recent ratings as the current rating, even if the player has not competed for a very long time

Appendix C
Glicko System Overview

Historical Information

- Developed by Glickman in 1999
- Developed as an extension of the Elo system. One may think of the Elo system as a special case of the Glicko system because it not only computes the player's rating but incorporates the reliability of the player's rating. This reliability is called *rating deviation* (*RD*).

Comparisons

- Two-player comparisons

Statistical Information

- First model with solid statistical foundation. It uses a Bayesian ranking system.
- Assumes that player's skill distribution follows a Gaussian distribution
- Applies the Bradley-Terry model
- The rating update for each player can be computed after each rating period. For better estimates, the number of games in each rating period should be between five and 10 games.

Advantages

- Attempts to improve the parameter estimates by incorporating the *RD*

Limitations

- It may not capture the true change in skills for players who compete frequently. This is because the *RD* is small for players who compete often. As a result, the rating for the players may not change accurately.

Other Information

- A Glicko-2 model adds a rating volatility to rating and *RD*. Rating volatility is the degree of expected fluctuation in player's rating.

Appendix D
Edo System Overview

Historical Information

- Developed by Rod Edwards
- Treats the same player at two different years as two different players

Comparisons

- Two-player comparisons

Statistical Information

- It is based on Bradley-Terry model. It provides variance of the player's skill.
- An adjustment is made to maintain the rating with a mean of 1,500 and a standard deviation of 300. The adjustment is made because more players at the lower end of the rating were included at the end of the 19th century.

Advantages

- It is claimed to estimate isolated players better than Glicko (Edwards, n.d.).

Limitations

- It is not a full Bayesian model, and it does not provide a posterior distribution.
- It does not model draws.
- It provides ratings only until 1910. The Edo system was developed in 2004, and it used old data for the purpose of rating.

Appendix E
TrueSkills System Overview

Historical Information

- Developed by Microsoft Research in 2007
- Adopted by Xbox Live game, Microsoft's Bing search engine, and Internet information multilabel classification
- A player's skill and performance are updated after each game.
- It is currently widely used.

Comparisons

- Two-player, two-player teams, multiplayer, and multiteam comparisons

Statistical Information

- Model building is flexible and is not limited to a two-player scenario.
- Ranking system also matches players and/or teams of players with similar skills so that gaming experience is more competitive and exciting.
- Gaussian distribution is assumed for each player's skill and performance. Each team's performance is the sum of its team members' performance. Draws are allowed in the system, and the margin of draw can be adjusted according to different ranking needs.
- Parameter estimation uses Bayesian approximation, factor graphs, and a sum-product algorithm. Bayesian approximation allows instantaneous ranking updates.
- Players are given a prior skill and the skill is updated after each game.

Advantages

- Model building is flexible and estimation is instantaneous. It also saves estimation resources—time and computing resources.
- Application of the TrueSkill variant model is popular and useful.
- It is reported that estimation of TrueSkill is more precise than that of Elo.

Limitations

- Bayesian approximation is a compromise among estimation precision, speed, and resources.
- The system will need some initial infrastructure building.