The Statistician (2001)
50, *Part* 2, *pp.* 191-208

Focus on Sport

New table-tennis rating system

David J. Marcus
Somerville, USA

[Received January 2000. Revised December 2000]

Summary. We present a new system for rating the playing strength of table-tennis players. The system is based on Bayesian principles and is designed to handle a large changing population of players, where some players play frequently whereas other players play infrequently. The system takes into account the length of time since a player last played a tournament. When processing matches in a single tournament, the system takes into account how a player's opponents did in the same tournament. The system has been tested by processing data from 5½ years of tournaments (15549 players and 330079 matches). The system could be adapted to other sports that involve head-to-head competition.

Keywords: Approximate Bayesian estimation; Dynamic paired comparisons; Sports rankings; Sports ratings; Table-tennis; Time-dependent abilities

1. Introduction

USA Table Tennis (USATT), the national governing body for table-tennis in the USA, has had a rating system since the 1970s. The system is similar to the chess rating system, which is undoubtedly the best-known sports rating system. However, the formulae are different and have changed several times over the years. We present a new table-tennis rating system that is designed to handle the characteristic features of US table-tennis tournaments. The system would also be suitable, perhaps with minor modifications, for other similar sports.

Table-tennis in the USA has a large changing population of tournament players that compete on an irregular schedule. Some players play a dozen tournaments a year, whereas others play less than one tournament a year. Only matches in USATT-sanctioned tournaments count for ratings. There is an average of 243 tournaments per year. The tournaments range in size from three players and three matches to 785 players and 6873 matches. The median is 50 players and 180 matches. Over the 5½ year period from January 1st, 1994, to June 30th, 1999, a total of 15549 different players played in tournaments, with each player playing in an average of 5.6 tournaments. Over the same period, there have been 330079 matches. Each match involves two players, so this is an average of 7.6 matches per player per tournament. Players tend to play other players who are at a similar playing level, but a significant number of matches are between players of quite different levels. The range of levels is quite large, ranging from the best players in the world to children who can barely see over the table.

Address for correspondence: David J. Marcus, 25 Beacon St. Apt. 16, Somerville, MA 02143-4336, USA.

© Royal Statistical Society 0039-0526/01/50001

There are several features of the problem of rating table-tennis players that make it fall naturally into the Bayesian framework. At any time, there is a wide range in how many matches players have played. Thus, it is natural to treat a player's level as uncertain. New results come in every week, so it is natural to process data by updating priors.

The new table-tennis rating system (NTTRS) (i.e. the system in this paper) uses a fully Bayesian model. The model is straightforward. New players are assigned a normal prior. Players who have competed before have a random walk added to their playing strength to reflect the passage of time. The probability that one player defeats another is a function of the difference in their playing strengths. Match results are processed by conditioning on the result (win or loss). Because of the large dimensionality of the problem, some approximations must be made to produce a workable algorithm. The NTTRS implements a non-iterative algorithm that approximates the posterior mean.

2. Sports ratings

A *rating system* estimates the playing strength of each player. In contrast, a *ranking system* only produces an ordinal list of players ordered by playing strength. If the values are going to be updated as new results come in, then ratings are preferable, since they maintain more information about the player's level. An obvious way to produce rankings from ratings is to sort the players by rating.

Many sports have rankings or ratings. Most are fairly simple, involving only weighting and averaging. Stefani (1997) and Bennett (1998), section 12.3, have surveyed the systems used by various sports.

The chess rating system is the best-known rating or ranking system. It uses a probability model, but it does not use conditioning to process new results. See Batchelder *et al.* (1992), Elo (1978), Glickman (1999a), Glickman and Jones (1999), Joe (1990), Sadovskiĭ and Sadovskiĭ (1993), and Strauss and Arnold (1987) for more information and related systems.

The American Go Association rating system maximizes the posterior of a Bayesian model. See http://ourworld.compuserve.com/homepages/accelrat.

Glickman (1999b) and Farhmeir and Tutz (1994) use Bayesian models that take time into account. Glickman used an approximate Bayesian updating algorithm. Section 11 compares an algorithm similar to Glickman's to the approximate algorithm that the NTTRS uses. Farhmeir and Tutz's algorithm is a state space smoothing algorithm. The NTTRS algorithm is neither a state space algorithm nor a smoothing algorithm.

Sports ratings have many connections to non-sports statistical problems. See the references in Batchelder *et al.* (1992), Farhmeir and Tutz (1994), Glickman (1999b), Joe (1990), and Stob (1984). Stob (1984) is a good introduction to some of the mathematics behind sports ratings.

3. Current system

This section describes the current table-tennis rating system. The current system is of interest because it is the official system of USATT and because the NTTRS was developed by using data that came from the current system.

The current system has two parts. The first part is the rating chart, shown in Table 1. To understand how this is used, consider an example: suppose that a player rated 1800 plays a player rated 1900. Subtract 1800 from 1900 to obtain 100. This is the *rating difference* between the two players. Find the row of the table that corresponds to this difference, i.e. the row '88–112'.

Suppose that the player rated 1900 won the match. Find the value in the 'expected result' column, i.e. 4. The winner of the match receives this many rating points, and the loser loses this many. Thus,

Table 1. Rating chart

Rating difference	*Expected result*	*Upset*
0–12	8	8
13–37	7	10
38–62	6	13
63–87	5	16
88–112	4	20
113–137	3	25
138–162	2	30
163–187	2	35
188–212	1	40
213–237	1	45
238–∞	0	50

the 1800 player's new rating would be 1796, and the 1900 player's new rating would be 1904. If the 1800 player won the match, then use the value in the 'upset' column instead, i.e. 20. In this case, the 1800 player's new rating would be 1820, and the 1900 player's new rating would be 1880.

By itself, the rating chart is a zero-sum system. We may think of it as a correction scheme (similar to the update formula in a state space system) or as a reward system (as described in Batchelder *et al.* (1992)). The numbers in the rating chart were basically just made up by the people who developed the current system. In fact, the original version had nine lines, so the numbers in the expected result column were simply the numbers from 0 to 8. The rating chart also functions as a filter to detect those players whose ratings need adjustment, as we shall see below.

The current system processes tournaments in the order that they are played. Each tournament is processed as an independent unit.

The system assigns an *initial rating* to each player in the tournament. If the player has played in a tournament before, the initial rating is set equal to the rating that the player had after the player's last tournament. Otherwise, the initial rating is 0, which is just an indicator that the player is unrated.

The current system makes three passes through the data from the tournament. Each pass produces a set of ratings for all the players.

For players who already had a rating (from playing in a previous tournament), the first-pass rating is just the initial rating. For unrated players, the system comes up with two suggested ratings. These are based on the player's performance against rated players in the tournament. The USATT Ratings Coordinator can select one of these two suggested ratings to be the first-pass rating or can make up a different rating by including subjective information (e.g. suggested ratings provided by tournament directors).

Table-tennis matches are either the best of three games or the best of five games with each game being to 21 points. When constructing the suggested ratings for unrated players, the system takes into account the scores (i.e. games and points) in each of the unrated player's matches. This is done as a sort of last resort, since often there is little other information on which to base a rating. Section 4 discusses the use of scores in a table-tennis rating system and whether it is a good idea.

In the second pass, the system applies the rating chart to all matches (using the first-pass ratings for all players). The resulting ratings are the *second-pass ratings*.

The system now determines *adjusted ratings* for each player:

(a) if the player gained fewer than 51 rating points in the second pass, the adjusted rating is equal to the first-pass rating;
(b) if the player gained between 51 and 75 rating points in the second pass, the adjusted rating is

Table 2. Summary report from current system

Name	*Initial rating*	*Adjusted rating*	*Matches played*	*Point change*	*New rating*
Aziz, Qassim	2180	2180	8	2	2182
Babb, Kim	0	1185	2	0	1185
Bacon, Dale	1565	1616	29	17	1633
Bahlman, Lee	2052	2052	26	19	2071
Baier, Owen	1598	1598	20	54	1652
Baikov, Dmitri	2314	2314	16	−55	2259
Bailey, Christopher T.	1654	1730	26	10	1740
Baksh, Raymond	0	1688	10	−3	1685

Table 3. Individual report from current system

Wins			*Losses*		
Point change	*Adjusted rating*	*Opponent*	*Point change*	*Adjusted rating*	*Opponent*
Aziz, Qassim: adjusted rating, 2180; new rating, 2182; point change, 2					
7	2165	Hinse, Pierre-Luc	-4	2276	Pandit, Sharad
4	2086	Yeung, Donne	-2	2327	Gagnon, Jean-Philippe
0	1799	Yee, Mario	-2	2325	Paulin, Karl
0	1791	Wei, Samuel	-1	2405	Therien, Xavier

equal to the second-pass rating;

(c) if the player gained more than 75 rating points in the second pass, the adjusted rating is set to a modified version of the second-pass rating. Typically, the modification is to average the second-pass rating with the average of the ratings of the highest rated opponent whom the player beat and the lowest rated opponent whom the player lost to. The USATT Ratings Coordinator can modify the adjusted ratings of these players (by using her judgment).

In the third (and final) pass, the system applies the rating chart to all matches (using the adjusted ratings for all players). The resulting ratings are the *third-pass ratings* and are also the new, post-tournament ratings for the players.

For each tournament, USATT generates two reports: a summary report that shows the rating points that each player gained or lost for the tournament and an individual report that shows the rating points that each player gained or lost for each match. Table 2 shows an excerpt from a summary report. The value in the 'point change' column is the new rating minus the adjusted rating, i.e. the points gained in the third pass. Most large rating point changes come from the adjustments, so the point change column can be quite misleading, but this is the way USATT does it. For example, Bailey actually gained $1740 - 1654 = 86$ rating points for the tournament, but his point change is shown as 10. Table 3 shows an excerpt from an individual report.

4. Using scores

This section discusses whether a rating system should use *scores*, i.e. the number of points and games that each player scored in a match. The current system only uses scores to assist in rating unrated players, i.e. players who are playing in their first tournament. The NTTRS does not use scores at all.

Clearly, if player P defeats player Q deuce (e.g. 22-20) in the fifth game, then we believe that the two players are much closer in level than if P wins three straight with each game under 10. However, there are quite a few practical considerations to consider before we decide that using scores in a rating system is a good idea.

Scores are not reliably recorded. Almost all matches played in US tournaments do not have umpires. Thus, the scores must be recorded by the players. The winner of the match is responsible for turning in the match card to the control desk. Often the score does not get written down until the control desk points out that the player is trying to turn in a match card without recording the score. Many scores get recorded as 21-15, 21-15 because the player can't remember what the score was but must write down something before he or she can turn in the match card.

Most events are either single-elimination or round robin groups with the winner of the group advancing to a single-elimination bracket. For the vast majority of matches, it is only relevant (from the tournament's perspective) who won the match, not what the score was. Thus, there is little incentive for players to work hard to make the score as lopsided as possible. The range of playing levels at a tournament is very large. Although tournaments are structured so that players mostly play other players of a similar level, there are many matches between players of significantly different levels. A fairly close score can mean that the players were closely matched, or it can mean that the better player was able to keep a small lead throughout the entire match and thus had no fear of being upset.

Players (and fans) think there is a big difference between winning deuce in the fifth game and losing deuce in the fifth game, even though the difference is only two points. This suggests that we think very carefully before deciding that the score is comparable in importance with the result.

Playing styles and tactics can affect the score. The fact that matches consist of several games is also relevant. Players may lose a game because their opponent has an unusual style but can then adjust their play and win the match. Or, a player may win the first two games easily, lose concentration and the third game, but then finish off the match. Thus, losing a game to an opponent does not necessarily mean the opponent is of a comparable playing level.

There is one situation in a tournament where the score is relevant to which player advances, and that is in the tie-break rules for round robin groups. However, because ties are rare, few players worry about this until the matches are completed. Thus, the existence of the tie-break rules is not enough to make players worry about the score.

If the rating system used scores, this would drastically affect how players view tournaments. Currently, all matches count for ratings, but this only means that players are concerned with winning their matches. (Ratings are an important attraction for players to play tournaments. Even if a player is unlikely to win money or a trophy, he or she can still win rating points.) If the rating system used scores, then players would become very conscious of how many points they were giving up in a win, or scoring in a loss. Presumably, if scores were used, a close win would not affect the players' ratings as much as a rout. While this would undoubtedly make the loser happy, it is unlikely that it would make the winner happy. Overall, it is quite likely that tournaments would be seriously, and adversely, affected.

Currently, there is a small, but significant, problem with players *dumping*, i.e. intentionally losing matches. Most tournaments in the USA run rating events. These are events where only players who have ratings below some cut-off are allowed to enter. For example, a tournament might have the following events: Under 1000 Singles, Under 1100 Singles, …, Under 2200 Singles, and Open Singles. (Note that a rating event is not the same as an event that counts for ratings. Almost all events at USATT-sanctioned tournaments count for ratings, i.e. the results affect the players' ratings.) At major tournaments, such as the US Open and the US National Championships, there is significant prize money in many of the rating events. Thus, there are players who lose intentionally in tournaments before to the major tournaments so that they can reduce their ratings sufficiently low to be eligible for rating events in which they have a better chance of winning. There are also cases of parents dumping to their children and top players dumping to their students. The top player may be older and not as concerned about his rating or may feel that he will be able to raise his rating by

beating other players. By losing to his student, his student can receive a rating boost and thus better seeding in events or possibly receive sponsorship.

If scores counted for ratings, it is very likely that many players would start fiddling with the score before turning in the match card. Losers in a match would feel that they had to walk with the winner to the control desk to ensure the winner did not change the score. Players would become annoyed at the tournament staff if any data entry errors were made. It would be a major headache.

Most data entry for the current system is done by USATT, i.e. tournament directors submit paper copies of the results, and the USATT Ratings Coordinator types the results into her computer. Eventually, we expect most tournament directors to submit digital results, but this will not be happening soon. It would significantly increase the work for USATT if scores had to be entered for every match.

The NTTRS appears to work perfectly well without considering scores. This is another reason not to use scores.

5. Model

This section presents the fully Bayesian model that is the foundation of the NTTRS. Section 7 explains the algorithm that the NTTRS uses.

Let $N(\mu, \Sigma)$ denote a normal law with mean μ and variance Σ. Each player has a true playing strength. Let

$$\alpha := 0.0148540595817432.$$

Let the function π be defined by

$$\pi(s) := \frac{1}{1 + \exp(\alpha s)}.$$

The probability that a player with playing strength s will lose to a player with playing strength t is $\pi(s - t)$. Call π the *probability-of-loss function*, and call α the *probability-of-loss parameter*. The probability-of-loss parameter determines the scale, i.e. what 100 rating points mean in terms of ability. Notice that the probability that one player defeats another only depends on the difference in their playing strengths. Two players who are 100 rating points apart are reasonably competitive (the probability that the weaker player wins is 0.18), whereas two players who are 200 rating points apart are at different levels of expertise (the probability that the weaker player wins is 0.05).

The functional form of the probability-of-loss function is the same as that in the Bradley-Terry model for paired comparisons (Bradley and Terry, 1952). This model was first used by Zermelo (1929).

Because we do not know the true playing strength of a player, it is modeled as a real random variable (thus following the Bayesian paradigm). The law (probability measure) of this random variable is referred to as the *player's law*. We need to know how to choose the law for a player who is playing in his or her first tournament (i.e. how to choose the prior). We also need to know how this law changes in time, i.e. between tournaments. Once we know these two items, the stochastic model is fully specified. Since we are following the Bayesian paradigm, match results are processed by conditioning on the result of each match.

The law $N(1400, 450^2)$ is assigned to players entering their first tournament and is called the *unrated prior*. Many players playing in their first tournament have been playing with their friends in their basement. These players often turn out to be at the 800–1200 level. Young children can be much lower. At the other extreme are world class players from abroad (e.g., the world champion) who come to the USA to play in the US Open. These players may be at the 3000 level.

The temporal evolution of a player's law is modelled by adding a random walk to the player's playing strength. Suppose that X is the player's playing strength after the player's last tournament.

Suppose that D days have elapsed since then. Let Y have law $N(0, 70^2 D/365)$. Then the model is that the player's playing strength is now $X + Y$. So, loosely speaking, the uncertainty in a player's playing strength increases at the rate of 70 rating points per year (actually, it is the variance that goes up linearly). The process of converting X to $X + Y$ is called the *temporal update*.

If the player's law did not change with time, then this would say that the player's playing strength does not change with time. The effect of this would be that old results would have equal weight with recent results. Thus, the temporal evolution component of the model effectively deweights results as they become old. It also controls how quickly the system responds, like a smoothing parameter.

There is one more piece to the model, which is really an initialization piece. The data set starts on January 1st, 1994. The law $N(1900, 600^2)$ is assigned to those players who had a rating under the current system on January 1st, 1994. It is natural to use a prior for the players who had a rating on January 1st, 1994, that is different from the unrated prior. Many of the players who had ratings on January 1st, 1994, were players that had been playing tournaments for many years, and thus many of them would be expected to be more proficient than new players. In other words, the two populations are different: one is the population of tournament players (at a specific time in the past), whereas the other is the population of first-time tournament players.

6. Model development

The model was developed using statistical analysis of data from the current system, experimentation and testing of various candidate models on historical data, and subjective input from tournament directors and players with many years of experience in the sport.

The value of the probability-of-loss parameter α came from a fit to data from the current system. Since α only determines the rating scale, and since testing showed that the value from the fit reproduced the scale of the current system, there was no reason to change, or round off, the value of α.

The unrated prior came from calculating the mean and variance of the ratings that the current system gave to unrated players. This was followed by testing to make sure that the NTTRS produced sensible ratings for tournaments like the US Open where many unrated players are world class.

The temporal update came from subjective input. This was followed by experimentation with three values: 30, 70, and 110 rating points per year. The value of 30 was clearly too small. Both 70 and 110 produced reasonable results, but 70 produced smoother results while still allowing the NTTRS to track rapidly improving players.

The prior for players who had a rating on January 1st, 1994, came from data from the current system. However, this was then modified (mean and variance both increased) so that the origin of the rating scale in the NTTRS (as of the present) would be closer to that of the current system.

7. Algorithm

With thousands of players and hundreds of thousands of matches, it is not obvious how actually to calculate anything from the model in Section 5. This section describes the algorithm that the NTTRS uses. The algorithm is not iterative (in contrast with, for example, finding the posterior mode via an iterative optimization algorithm). Section 8 gives explicit formulae for the algorithm.

Instead of letting playing strength be a real number, each player's playing strength is assumed to be in the set $\Omega := \{0, 10, 20, \ldots, 3600\}$. Since the model uses normal laws, we need to use discrete versions of the normal law. Any reasonable approximation would probably be adequate, but to be precise the mass of the normal law is assigned to the point in Ω that is closest. For example, we

assign to the point 0 the probability that the normal law assigns to the interval $(-\infty, 5]$ and assign to the point 10 the probability that the normal law assigns to the interval $[5, 15]$.

There is, between tournaments, no attempt to keep any joint information; only the marginal law for each player is saved.

As a tournament is processed, players have several laws. At the beginning of a tournament, a player has an *initial law*. At the end of a tournament, a player has a *final law*. To obtain the initial law, the temporal update (which is now just a discrete convolution) is applied to the player's final law from after the player's last tournament.

The remaining question is how to condition on the match results from a single tournament. Taking the joint law of all the players and conditioning on the match results is impractical. We could process each of a player's matches by using the opponent's initial law, but this would discard too much information. Many players play infrequently. Assuming that the playing level of such players is unchanged since their last tournament is too extreme. (Section 11 gives some experimental evidence for this assertion.)

Suppose that we want to update player P's law to take into account P's matches. Let Q be one of P's opponents. Define Q's *adjusted law (for P)* to be the law that we obtain by conditioning Q's law on all the results of Q's matches, except those matches with P, and using, for each of Q's opponents, the opponent's initial law. To process P's matches, condition on the results of P's matches using each opponent's adjusted law.

If two players play more than one match with each other at the same tournament, then their results with each other must be processed as a unit, rather than one at a time.

8. Formulae

This section gives formulae for the algorithm. To process the results of a tournament, each player's initial law $\mathcal{B}$ is calculated as follows.

(a) If the player has not played in a previous tournament, then set $\mathcal{B}$ to be a discrete version of $N(1400, 450^2)$. However, if the player had a rating in the current system on January 1st, 1994, then set $\mathcal{B}$ to be a discrete version of $N(1900, 600^2)$.

(b) If the player has played in a previous tournament, then retrieve the player's final law (after the player's last tournament) from the database. Call this law $\mathcal{F}$. Let D be the number of days since the player's last tournament. Let the law $\mathcal{T}$ be a discrete version of $N(0, 70^2 D/365)$ on $\{-3600, -3590, \ldots, 3600\}$. Then set $\mathcal{B}$ so that it satisfies

$$\mathcal{B}(r) = \begin{cases} \sum_{x\in\Omega} \sum_{y\le -x} \mathcal{T}(y)\mathcal{F}(x), & r = 0, \\ \sum_{x\in\Omega} \mathcal{T}(r-x)\mathcal{F}(x), & r \in \Omega \setminus \{0, 3600\}, \\ \sum_{x\in\Omega} \sum_{y\ge 3600-x} \mathcal{T}(y)\mathcal{F}(x), & r = 3600. \end{cases}$$

This is basically just a discrete convolution, but we have modified the values of $\mathcal{B}(0)$ and $\mathcal{B}(3600)$ to include the probability that would otherwise leak out of the ends of Ω (because of the discretization).

Before we can describe how to process the matches, we need to know how to update a law for the match results between two players. Suppose that we have two players P and Q and we want to update P's law by conditioning on the results between P and Q. Let $\mathcal{U}$ be the resulting updated law for P. Let W be the number of matches that P won, and let L be the number of matches that P

lost (when P and Q played each other). Let $\mathcal{L}_P$ be P's law, and let $\mathcal{L}_Q$ be Q's law. Let U be the real-valued function on Ω defined by

$$U(p) := \sum_{q\in\Omega} \pi(q-p)^W \pi(p-q)^L \mathcal{L}_Q(q)\mathcal{L}_P(p), \qquad p \in \Omega.$$

Then $\mathcal{U}$ satisfies

$$\mathcal{U}(p) = \frac{U(p)}{\sum_{q\in\Omega} U(q)}, \qquad p \in \Omega.$$

We are now ready to calculate the final law for each player. The process is as follows. Pick a player P. Start with the player's initial law. Consider the player's first opponent Q_1. Calculate Q_1's adjusted law (for P). Using Q_1's adjusted law, update P's law for P's matches with Q_1. Then take the law that results and (using Q_2's adjusted law) update it for P's matches with P's next opponent Q_2. Keep going until all P's opponents Q_i have been processed. This gives P's final law. Now do the same for each of the other players. This completes the algorithm.

9. Speeding up the algorithm

There are a few tricks that will speed up the algorithm. Powers of the function π may be precalculated. The laws $N(0, 70^2 D/365)$ may be saved, once calculated, since the same values of D will often recur.

A player's adjusted law can be calculated backwards. First, a law $\mathcal{F}$ is calculated for player Q and updated for all of Q's matches (using initial laws for each of Q's opponents). Now, suppose that we are processing player P, and we need Q's adjusted law (for P). Let W be the number of matches that P won, and let L be the number of matches that P lost (when P and Q played each other). Let $\mathcal{L}_P$ be P's initial law. Let V be the real-valued function on Ω defined by

$$V(q) := \frac{\mathcal{F}(q)}{\sum_{p\in\Omega} \pi(q-p)^W \pi(p-q)^L \mathcal{L}_P(p)}, \qquad q \in \Omega.$$

Then Q's adjusted law $\mathcal{V}$ satisfies

$$\mathcal{V}(q) = \frac{V(q)}{\sum_{p\in\Omega} V(p)}, \qquad q \in \Omega.$$

The reason that this technique is useful is that $\mathcal{F}$ is calculated only once and then reused for each of Q's opponents. Calculating the adjusted law in this way is only worthwhile if Q has at least four opponents. To see this, suppose that Q has four opponents: P_i, $i = 1, 2, 3, 4$. To calculate Q's updated law for P_1 requires starting with Q's initial law and updating it in turn for the matches with P_2, P_3, and P_4, i.e. doing 3 updates. If we calculate all Q's updated laws similarly, then we shall need to do 12 updates. Suppose that we use the backward algorithm. First we calculate $\mathcal{F}$; this takes four updates. Starting with $\mathcal{F}$, we only need to do one update (actually an 'un-update') to calculate each of Q's updated laws. This makes a total of eight updates.

In practice, the support of a law is usually a proper subset of Ω. Some disk space can be saved by storing only the values within an interval containing the support.

Using these tricks, it takes a personal computer about 8 hours to process the 1336 tournaments from January 1st, 1994, to June 30th, 1999. The main database containing the laws is around 50 Mbytes.

10. Normality

Are all the laws normal? All the priors are normal, and the random walk in the temporal update is normal. However, the laws do not stay normal because conditioning on a match result will not produce a normal posterior.

Some players will have laws that are skew. In particular, players who win (or lose) all their matches in their first tournament can come out with posterior laws that are skew. This is fairly common since players in their first tournaments do not have ratings and thus are sometimes placed in too low (or high) an event. However, most laws that the system produces look bell shaped, and so we might consider fitting normal laws to the posteriors. Certainly, only saving two numbers for each player would save a considerable amount of disk space. For table-tennis, the disk space requirements are not excessive, but for more popular sports, this could be a consideration.

Using discrete laws is tantamount to a decision to use Riemann sums to do integrals. We could consider using a different integration scheme, perhaps Gauss-Hermite. Another alternative is to use the formulae in Glickman (1999b) to calculate first the means and variances of the adjusted laws, and then use Glickman's formulae with these means and variances to calculate the mean and variance of the final law.

11. Tournament surgery

The algorithm may be viewed as modifying the tournament and then processing the modified tournament. For any player P, let P' be P's *twin*, i.e. another player who has the same initial law as P. The example in Fig. 1 illustrates the following discussion.

A tournament may be thought of as a graph with each player a node and each match an edge connecting two players. Consider a particular node (e.g., player P in Fig. 1). All that player's opponents will be one level below (the *first level*). The opponents of the opponents will be one level below that (the *second level*). Cut all the edges that extend down from the second level (e.g., match m_5). Discard all the edges that connect two nodes at the second level (e.g., m_4). For each edge connecting two nodes at the first level (e.g., m_1), cut the edge and insert two new nodes on the newly created ends (e.g., Q'_1, Q'_2). The new nodes are twins of the two original nodes. (In Fig. 1, new nodes are drawn using squares.) If a node at the second level connects to two nodes at the first level (e.g., Q_4), turn the node at the second level into twins (e.g., Q_4 and Q'_4) and connect each of the nodes at the first level to one and only one of the twins.

The algorithm uses this resulting graph instead of the original one. For this graph, the algorithm calculates the true Bayesian posterior (for the discrete model) for the player (node) selected. Thus, if the tournament was such that for some player we did not need to do any cutting, etc., then the algorithm is exact, rather than only approximate, for that player. For most players in most tournaments, the algorithm will only be approximate.

Notice that it does not matter in what order the algorithm processes matches. This is because the algorithm is calculating a Bayesian posterior; it just may not be the Bayesian posterior that corresponds to the real tournament. In other words, the system approximates the tournament and then calculates (using discrete models) the exact posterior for the approximate tournament. (To be precise, it calculates the marginals of the posterior, not the joint posterior of all the players.) The trick is that the approximating tournament depends on the player whose final law the system is calculating.

A different approach to calculating the final law is to use the initial law for each opponent. Call this the *IL* algorithm. For example, Glickman (1999b) uses this approach. This approach can also be viewed as using an approximate tournament.

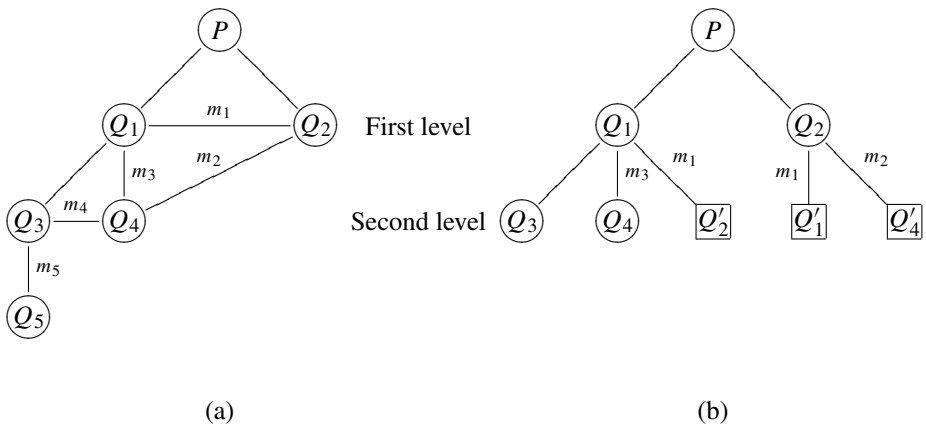

Fig. 1. Tournament surgery: (a) before; (b) after

Table 4. Tournaments used for $A \sim B \sim C \sim D$

Player	*Tournaments used by the following methods:*	
	NTTRS	*IL*
A	$A \sim B \sim C$	$A \sim B$
B	$A \sim B \sim C \sim D$	$A \sim B \sim C$
C	$A \sim B \sim C \sim D$	$B \sim C \sim D$
D	$B \sim C \sim D$	$C \sim D$

Let $A \sim B$ denote that player A played one match with player B. Say that an algorithm is *exact* for a player if the law that the algorithm calculates for the player is the same as we would obtain by calculating the marginal for the player from the full joint posterior (in both cases using the discrete models).

Suppose $A \sim B \sim C \sim D$. Table 4 shows the tournament that is used for each player. The NTTRS algorithm is exact for B and C, but not A or D. The IL algorithm is not exact for any of the players.

Suppose $A \sim B \sim C \sim A$. Table 5 shows the tournament that is used for each player. Neither algorithm is exact for any of the players.

In general, the NTTRS algorithm is exact for a player as long as each opponent of the player's opponents only plays one player. To be precise, let

$$\mathcal{O}(P) := \{\, Q \mid P \sim Q \,\}.$$

If $\mathcal{A}$ is a set of players, then let

$$\mathcal{O}(\mathcal{A}) := \bigcup_{P \in \mathcal{A}} \mathcal{O}(P).$$

Table 5. Tournaments used for $A \sim B \sim C \sim A$

Player	*Tournaments used by the following methods:*	
	NTTRS	*IL*
A	$B' \sim C \sim A \sim B \sim C'$	$C \sim A \sim B$
B	$C' \sim A \sim B \sim C \sim A'$	$A \sim B \sim C$
C	$A' \sim B \sim C \sim A \sim B'$	$B \sim C \sim A$

Table 6. Statistics of the log-posterior for 1336 tournaments

Statistic	$\log(f(\nu)/f(\mu_0))$	$\log(f(\nu)/f(\iota))$
Average	213.82	5.71
Standard deviation	552.34	27.63
Median	135.07	3.56
Minimum	−23.20	−116.41
Maximum	12956.04	288.12

Table 7. Log-posterior for large tournaments

Tournament	*Players*	*Matches*	$\log(f(\nu)/f(\mu_0))$	$\log(f(\nu)/f(\iota))$
1994 Open	672	2849	4894.46	260.20
1994 Teams	769	6873	12956.04	176.89
1994 Nationals	601	3291	2236.83	102.70
1995 Open	581	2611	2589.97	119.67
1995 Teams	620	5542	6084.93	185.42
1995 Nationals	660	3387	1519.41	100.02
1996 Open	670	2998	3298.33	288.12
1996 Teams	678	6255	5931.15	30.12
1996 Nationals	613	3084	1071.68	19.62
1997 Open	785	3259	4382.16	−116.41
1997 Teams	699	6527	7111.35	2.83
1997 Nationals	650	3194	1258.63	126.75
1998 Open	524	2328	1591.71	65.55
1998 Teams	629	5596	6184.52	19.45
1998 Nationals	592	3001	1122.62	20.42

Then the algorithm is exact for P if for all Q in

$$\mathcal{O}(\mathcal{O}(P)) \setminus \{P\}$$

we have

$$|\mathcal{O}(Q)| = 1.$$

To get some idea of how well the algorithms approximate the posterior mean, we can evaluate the log-posterior at the various estimates. Let f be the density of the posterior. Let μ_0 be the prior mean, i.e. the vector for all the players in a tournament. Let ν be the posterior mean that the NTTRS calculates, and let ι be the posterior mean that the IL algorithm calculates. Table 6 gives some statistics for $\log(f(\nu)/f(\mu_0))$ and $\log(f(\nu)/f(\iota))$ for the 1336 tournaments. For this data set, the NTTRS algorithm comes closer, on average, to the posterior mode than does the IL algorithm.

Each year in the USA, there are three major tournaments: the US Open, US Nationals, and US Open Team Championship. Over the 5½ year period, these tournaments have had 524–785 players and 2328–6873 matches. (The next largest tournament in the USA had 359 players and 1107 matches.) Table 7 gives the log-posterior values for these tournaments. For 14 of the 15 tournaments, the NTTRS algorithm comes closer to the posterior mode than does the IL algorithm.

12. Examples

This section gives some numerical examples for the NTTRS algorithm. Let $A > B$ denote that player A defeated player B.

Table 8. $A > B$, and B is initially $N(2000, 60^2)$

A's mean	*B's final mean*	*B's final standard deviation*
2500	2000	60
2000	1977	56
1800	1953	58
1500	1947	60
1000	1946	60
500	1946	60

Table 9. $A > B > C$, and B is initially $N(1900, 100^2)$

A's mean	*A's standard deviation*	*C's mean*	*C's standard deviation*	*B's final mean*	*B's final standard deviation*
2000	0	1800	0	1900	78
1800	0	2000	0	1900	78
2000	60	1800	60	1900	84
1800	60	2000	60	1900	80

Consider two players A and B. Suppose $A > B$ and B's law is $N(2000, 60^2)$. Table 8 gives B's final mean and standard deviation for various initial means for A. In each case, A's initial (and final) standard deviation is 0. All final values are rounded to the nearest integer. The change in B's mean is monotonic. However, B's standard deviation initially drops, but then rises as the upset becomes implausible. Also, note that the change in B's mean is always less than B's initial standard deviation.

Suppose $A > B > C$ and player B is $N(1900, 100^2)$. Table 9 gives B's final mean and standard deviation (both rounded to the nearest integer) for various normal initial laws for A and C. When B's opponents have standard deviations of 0, it does not matter whether B beats the better opponent or the worse opponent: B's new rating is the same. (It is not difficult to check that the likelihoods are proportional in these two cases.) However, when B's opponents have positive standard deviations, it does matter. The result (which may be intuitively surprising) is that we are surer that B is rated 1900 if B has the less likely result of defeating the stronger opponent and losing to the weaker opponent. Note that the NTTRS algorithm is exact for B in this example, so the results cannot be explained by concluding that the NTTRS algorithm does not evaluate the model correctly.

The explanation is to think about all three players together. Since A defeated B, player A is probably stronger than we thought. Similarly, player C is weaker than we thought. If A is better than C, this implies that A and C are further apart than we originally thought. However, in the case where A is worse than C, this implies that A and C are closer together. Since B's results put him in between, B comes out with a smaller standard deviation in the latter case.

The standard deviation almost always decreases as matches are processed. However, here is an example where a result increases the standard deviation. Let player A be $N(1400, 450^2)$. Suppose player A defeats eight different $N(1200, 50^2)$ players. Then player A's mean will be 1744 and his standard deviation will be 282. Now suppose that A defeats another four players, each $N(1400, 450^2)$. Then A's new mean will be 1946 and his standard deviation will increase to 286. Note that in a model where everything is normal and linear (e.g., a state space model as used in a Kalman filter) it is impossible for the standard deviation to increase when data is processed.

Here is an example motivated by the following scenario. Suppose that players A, B, and C play a round robin and $A > B > C > A$. Player A is declared the winner via some tie-break procedure (e.g., games won or points scored) and advances to the next round where he loses to player D. Suppose that all four players start out as $N(1800, 50^2)$. Intuitively we might think that A should

Table 10. Round robin example

Player	*Results for the exact method*		*Results for the NTTRS method*		*Results for the IL method*	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
A	1787.30	43.72	1787.32	43.69	1787.27	43.78
B	1798.66	45.51	1798.77	45.47	1800.00	45.59
C	1798.66	45.51	1798.79	45.48	1800.00	45.59
D	1815.37	47.67	1815.37	47.67	1815.16	47.74

Table 11. Summary report for the new system†

Name	*Previous rating*	*Initial rating*	*Matches played*	*Point change*	*New rating*
Alban, Keith	2340 (34)	2340 (37)	4	−1	2339 (35)
Alderman, Diane	441 (198)	441 (201)	2	0	441 (201)
Alderman, Jerry Lewis	1614 (46)	1614 (58)	7	18	1632 (44)
Andrzejewska, Danuta	1655 (52)	1655 (59)	12	−8	1647 (45)
Azran, Hanan	757 (76)	757 (82)	7	1	758 (72)
Baumgartner, Peter	1574 (64)	1574 (83)	13	19	1593 (49)
Bavinger, Clay James	758 (114)	758 (134)	6	103	861 (89)
Bennett, Jermaine	0	1400 (450)	3	−171	1229 (253)
Bocanegra, Jose	1375 (55)	1375 (56)	12	80	1455 (42)

†Standard deviations are given in parentheses

come out with a mean at least as high as those of B and C, since A won the round robin group. However, the NTTRS only sees wins and losses (as discussed in Section 4) and so is unaware of the information that was used to break the tie. Also, A has a loss to D. (In a more realistic situation, D would presumably have other matches, since he reached the next round also, but, for this example, suppose that D's only match is his win over A.)

Table 10 gives the final means and standard deviations for each of the players, using the exact calculation, the NTTRS algorithm, and the IL algorithm. The exact calculation took over an hour of computer time whereas the algorithms each took under a second. We see that A's extra loss has caused his final mean to be less than those for the other members of his round robin group. The exact calculation sees the cycle among A, B, and C and gives B and C the same final mean and standard deviation. The NTTRS comes close but gives B and C slightly different means and standard deviations.

13. New system reports

For each tournament, the NTTRS generates two reports that are analogues of the reports that the current system generates. Table 11 shows an excerpt from a typical summary report from the NTTRS. For each rating the mean is given followed, in parentheses, by the standard deviation. The 'previous rating' column is the rating after the player's last tournament before this one. The 'initial rating' column is the rating of the player at the start of this tournament. For players who have played tournaments before, the mean in the two columns is the same whereas the standard deviation increases because of the temporal update. Unrated players have the unrated prior as their initial rating. The 'point change' column is the change in the mean from the initial rating to the new rating.

Table 12 shows an excerpt from a typical individual report from the NTTRS. Notice that an opponent's adjusted rating depends on the player. For example, Baumgartner beat Bocanegra, and Chan lost to Bocanegra. However, Bocanegra's adjusted rating is different in the two matches

Table 12. Individual report for the new system†

Wins			*Losses*		
Point change	*Adjusted rating*	*Opponent*	*Point change*	*Adjusted rating*	*Opponent*
Baumgartner, Peter; initial rating, 1574 (83); new rating, 1593 (49); point change, 19					
45	1685 (42)	Look, Raymond	−53	1455 (71)	Caplin, Stuart
12/2	1621 (48)	Mannarino, Mark	12/2	1621 (48)	Mannarino, Mark
14	1567 (67)	Mills, Nesly	−11	1666 (44)	Fong, George
17	1453 (44)	Bocanegra, Jose	−6	1704 (54)	Harley, Eddie
1	1289 (43)	Williams, Derrick		1934 (46)	Smith, Lynwood J.
	1130 (50)	Hackler Jr., Ted H.			
	1130 (50)	Hackler Jr., Ted H.			
	797 (193)	Gardner, Jim			
Chan, Pedro; initial rating, 1499 (79); new rating, 1446 (64); point change, −53					
4	1235 (39)	King, Hawathia	−47	1434 (44)	Bocanegra, Jose
2	1186 (37)	Bernstein, Joshua	−10	1595 (51)	Schoeman, Abel
	797 (194)	Gardner, Jim	−3	1688 (42)	Simpson, Daniese
				2026 (46)	Chu, Bin Hai

†Standard deviations are given in parentheses

(because one is for Baumgartner and one is for Chan, not because one match was a win and the other a loss).

To the left of each adjusted rating is the number of rating points the player gained for that match (i.e. the change in the mean). For example, Baumgartner gained 45 rating points from defeating Look. This value depends on the order in which the NTTRS processed the matches, but this order is not shown. So, the rating points gained per match is ill-defined. Despite this, it gives some clue about what the system is doing. (Note that the sum of the rating point changes is independent of the processing order.)

Multiple matches between the same two players are processed as a unit. For example, Baumgartner gained 12 rating points total from his two matches with Mannarino.

14. Comparison of new system with current system

This section lists some benefits of the NTTRS compared with the current system. Further explanation and discussion may be found in Marcus (1999).

(a) Ratings under the NTTRS do not fluctuate as much as under the current system. Ratings under the NTTRS fluctuate only when they should, i.e. when the player's playing level fluctuates.
(b) The NTTRS does not have a 51 rating point threshold where adjustments suddenly cut in. Under the NTTRS there will be no incentive for players to look for major upsets so that they can get the adjustment to kick in. Under the NTTRS there is no incentive to go into a large tournament underrated so a few good results will cause the player's rating to be adjusted and overshoot.
(c) The NTTRS treats players who have played many tournaments differently from players who have played only a few matches, i.e. it is more cautious in changing their ratings. The NTTRS treats players who have not played a tournament in years differently from players who have played a tournament recently, i.e. it is quicker to adjust their ratings.

(d) The NTTRS is fully automatic and objective, whereas the current system requires human assistance to handle the two tasks of assigning initial ratings and doing adjustments. Besides avoiding the occasional outright blunder that currently occurs, the NTTRS does a better job since these two tasks are, in reality, too difficult for a human to do either consistently or well.
(e) The NTTRS appears to be more accurate. Marcus (1999) gives examples of ratings from the current system that are wrong by 1500 rating points.
(f) The NTTRS can process tournaments as they become available rather than having to process in strict chronological order. This is because the NTTRS does not require human assistance, and thus the software implementation can be set up to go back and insert missing tournaments as they become available. Therefore, under the NTTRS, a tardy tournament director in Florida (for example) will not hold up the processing of tournaments in New York, as happens under the current system.
(g) The NTTRS keeps track of the accuracy of each player's rating. This will be helpful to tournament directors, both for determining eligibility to compete in events and, in special circumstances, seeding. It will also be of interest to players, allowing them to judge their opponents better.
(h) The NTTRS does not require scores whereas the current system uses scores for unrated players.

15. Extensions

This section discusses some possible extensions of the NTTRS.

15.1. Weighting important events

When calculating ratings or rankings, many sports weight matches by their importance. For example, major tournaments may count more than minor tournaments, or finals may count more than first-round matches. In table-tennis, we might wish to count the professional events at the US Open or at the US National Championships more than other events or other tournaments.

To incorporate such a feature into the NTTRS, it is natural to do so by using a different probability-of-loss function for those matches that should count more. Some experimentation would be necessary to determine a suitable probability-of-loss function.

15.2. Match format

In table-tennis, matches are the best of three games or the best of five games. Most matches are the best of three games. It would be quite reasonable to use a different probability-of-loss function for best-of-five matches than for best-of-three matches.

A first cut at constructing the probability-of-loss function for best-of-five matches would be to take the probability of winning a best-of-three match, to convert it to the equivalent probability of winning a point, and then to convert this to the probability of winning a best-of-five match. To go from the probability of winning a point to the probability of winning a match, each point is assumed independent. The calculation is then straightforward if we note that (ignoring the minor complication of deuce) a game is the best of 40 points (i.e. do not stop the game when one player reaches 21; play all 40 points). Deuce games are handled by using a recurrence formula. Similar calculations are carried out in Marcus (1985) and in Strauss and Arnold (1987).

Modelling each point as independent may be a poor model of the real world, so some testing and experimentation would be needed to come up with a suitable probability-of-loss function. However, the idealized calculation would provide some guidance.

15.3. Unrated prior for professionals

Only a couple of US tournaments each year draw international professional players. It would make sense to use a different unrated prior for these players. However, the NTTRS currently appears to be coming up with sensible post-tournament ratings for these players. Not too surprisingly, the match results for these players are much more influential than their priors. Thus, there does not seem to be a need for a different unrated prior for professional players.

15.4. Temporal update for juniors

It seems reasonable that juniors improve more rapidly (on average) than do other players. Thus, it might make sense to use a different temporal update for juniors than for other players. However, the NTTRS currently appears to be having no trouble in tracking the improvement of top juniors and other rapidly improving players. Thus, there does not seem to be a need for a different temporal update for juniors. A practical consideration is that USATT (somewhat surprisingly) does not have dates of birth for many players.

Back in the early 1980s, rapidly improving players were chronically underrated. In the mid 1980s, the official rating system was modified to be more reactive. This seems to have mostly fixed the problem with rapidly improving players although introducing excessive volatility.

15.5. Doubles

It would be very interesting to develop a system for rating the doubles strength of players. One approach is to rate doubles teams. However, players in the USA tend to play with many different doubles partners. Thus, many teams probably do not play enough matches to get reliable ratings. Another consideration is that it would probably be more interesting to the average player if there was a way to rate his or her individual doubles ability.

Almost all players who play doubles also play singles. There is a strong correlation between doubles ability and singles ability. With these two sketchy facts as motivation, here is an idea for a doubles system. Let the doubles strength of a player be the player's singles strength plus an independent term which could have a prior law of, say, $N(0, 100^2)$. Let the strength of a doubles team be the average of the doubles strengths of the two players and the singles ratings of the players be used to provide laws for the singles strength terms. The results of doubles matches would update the laws of the doubles strengths.

A problem with developing such a system is that USATT does not have any doubles data. This situation may change in the future as new software is developed for tournament directors to use.

16. Concluding remarks

The USATT Ratings Committee has recommended that USATT adopt the NTTRS. The decision on whether USATT will adopt the NTTRS now rests with the USATT Board of Directors. The USATT Ratings Committee believes that there is a reasonable chance that USATT will adopt the new system, eventually.

Acknowledgements

This work was supported by a grant from the US Olympic Committee's Sport Science and Technology Committee. Thanks go to Larry Rose (who developed the latest version of the current system) for his suggestions and assistance. Thanks go to Sean O'Neill for insisting that I undertake this project.

(Larry, Sean, and I currently constitute three-quarters of the voting members of the USATT Ratings Committee.) Thanks to Aaron Avery for pointing out that the adjusted laws could be calculated backwards (thus decreasing the run time by a third). And, thanks go to Jim Foltz, Albyn Jones, Mark Glickman, the referee, and the Joint Editor for their comments and suggestions.

References

Batchelder, W. H., Bershad, N. J., and Simpson, R. S. (1992) Dynamic paired-comparison scaling. *J. Math. Psychol.*, **36**, 185–212.

Bennett, J. (ed.) (1998) *Statistics in Sport*. London: Arnold.

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

Elo, A. E. (1978) *The Rating of Chessplayers, Past and Present*. New York: Arco

Farhmeir, L. and Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Am. Statist. Ass.*, **89**, 1438–1449.

Glickman, M. E. (1999a) Chess rating systems. *Am. Chess J.*, no. 3, 59–102.

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

Glickman, M. E. and Jones, A. C. (1999) Rating the chess rating system. *Chance*, **12**, 21–28.

Joe, H. (1990) Extended use of paired comparison models, with application to chess rankings. *Appl. Statist.*, **39**, 85–93.

Marcus, D. J. (1985) Probability of winning a game of racquetball. *SIAM Rev.*, **27**, 443–444.

Marcus, D. J. (1999) Benefits of new rating system. *Technical report*.

Sadovskiĭ, L. E. and Sadovskiĭ, A. L. (1993) *Mathematics and Sports*. Providence: American Mathematical Society.

Stefani, R. T. (1997) Survey of the major world sports rating systems. *J. Appl. Statist.*, **24**, 635–646.

Stob, M. (1984) A supplement to "A mathematician's guide to popular sports". *Am. Math. Monthly*, **91**, 277–282.

Strauss, D. and Arnold, B. C. (1987) The ratings of players in racquetball tournaments. *Appl. Statist.*, **36**, 163–173.

Zermelo, E. (1929) Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Warscheinlichkeitsrechnung. *Math. Zeit.*, **29**, 436–460.

The definitive version of this article is available at www.blackwell-synergy.com.