ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Models and Rating Systems for Head-to-Head Competition

Mark E. Glickman[1] and Albyn C. Jones[2]

[1] Department of Statistics, Harvard University, Cambridge, Massachusetts, USA; email: glickman@fas.harvard.edu

[2] Department of Mathematics, Reed College, Portland, Oregon, USA

Annu. Rev. Stat. Appl. 2025. 12:21.1–21.24

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040722-061813

Keywords

competitor strength, head-to-head competition, paired comparisons, playing strength

Abstract

One of the most important tasks in sports analytics is the development of binary response models for head-to-head game outcomes to estimate team and player strength. We discuss commonly used probability models for game outcomes, including the Bradley–Terry and Thurstone–Mosteller models, as well as extensions to ties as a third outcome and to the inclusion of a home-field advantage. We consider dynamic extensions to these models to account for the evolution of competitor strengths over time. Full likelihood-based analyses of these time-varying models can be simplified into rating systems, such as the Elo and Glicko rating systems. We present other modern rating systems, including popular methods for online gaming, and novel systems that have been implemented for online chess and Go. The discussion of the analytic methods are accompanied by examples of where these approaches have been implemented for various gaming organizations, as well as a detailed application to National Basketball Association game outcomes.

1. INTRODUCTION

In recent years, the field of sports analytics has overcome its initial perception as a mere niche interest, evolving into a full-fledged area of scientific inquiry and development within sports science and data science. This transformation is arguably attributed to the careful application of advanced statistical models, machine learning techniques, and data-driven strategies to enhance athletic performance, team dynamics, and overall game strategies. As a result, sports analytics not only has reshaped the way athletes train and compete, but also has provided a promising platform for researchers, coaches, and sports organizations to explore the complexities of sports in a more empirical and systematic manner. Recent articles that have reviewed the burgeoning landscape of sports analytics include those of Morgulev et al. (2018), Szymanski (2020), and Watanabe et al. (2021).

A substantial subdomain within sports analytics is the development and refinement of methods for estimating team or player abilities and constructing predictive models for game outcomes. The two goals are intrinsically linked; to make accurate predictions of game outcomes, one needs to make accurate inferences about the competitors involved in the games. A wide variety of approaches have been developed for different types of applications. Games with point scores have been modeled across various sports, including models for tennis outcomes (Spanias & Knottenbelt 2013), for National Football League (NFL) scores (Glickman & Stern 2017), and for the outcomes of low-scoring games like soccer (Scarf & Rangel Jr. 2017). In games that involve multiple simultaneous competitors, models for rank orderings have become increasingly popular, relying on models such as the Plackett–Luce model (Plackett 1975, Luce 1959). These approaches have also been extended to account for dynamic abilities, the incorporation of game-specific covariates, and a variety of other analytic features that are designed to improve game prediction and strength estimation.

Predictive modeling in sports analytics focuses to a large extent on binary outcome models in head-to-head games. Unlike point score or ranking models, binary outcome models simplify the prediction to the most fundamental question: Which team will win or lose? This approach, while seemingly reductive, has the potential to be more robust and therefore less biased than point score models by eliminating the possibility of using misspecified models for game scores. By distilling the prediction process to win-loss outcomes, these models offer a clear-cut perspective on game predictions, making them particularly appealing for their simplicity and direct application in betting markets and win probability analyses.

The focus of this article is on these binary outcome models, and rating systems that are founded on these models. We illustrate the application of these methods on National Basketball Association (NBA) game outcomes. The data consist of game outcomes played during the regular seasons from 2004–2005 to 2018–2019. Games played during the postseason playoffs, which often have different dynamics and coaching elements than regular season games, are not included in our analyses. We retain the final season (2018–2019), the last season before the COVID-19 pandemic, as a holdout set for predictions and develop the models and rating systems on game data through the 2017–2018 season. During the period from 2004–2005 to 2017–2018, the NBA consisted of 30 teams, each of whom played 82 games per season, resulting in a total of 1,230 games played each season. The 82 games are not distributed uniformly across opposing teams; pairs of teams compete anywhere between 2 to 4 times within a season depending on whether the opposing team is within the same division or conference. Thus, the game schedule within a season is imbalanced. In two seasons during this period, fewer than 1,230 games were played. The 2011–2012 season involved only 990 games due to a lockout in which the NBA owners and the players' union were unable to agree on a new collective bargaining agreement until midway through the season. Once the dispute was resolved, each team played only 66 games during the 2011–2012 season. In the 2012–2013 season, a game scheduled in April 2013 between the Boston Celtics and the Indiana Pacers to be played in Boston was cancelled in the wake of the Boston Marathon bombing, in part because of security concerns. The game was not rescheduled, and this resulted in a total of 1,229 regular season games played during the 2012–2013 season. Unlike other professional sports like NFL football or soccer, all NBA games are decisive. We use the binary game outcomes in our analyses of team strength and for making game predictions.

The organization of this article is as follows. We introduce the basic probability models in Section 2. Here we consider models not only for binary outcomes, but also for models in which a tie can be a third outcome. In Section 3, we review extensions of the basic models to permit competitor strengths that evolve over time. These models can be computationally intensive to fit and summarize, which motivates the development of rating systems for head-to-head competition. Popular rating systems are introduced in Section 4, with a special focus on the Elo (1978) and Glicko (Glickman 1999) rating systems. We review other modern rating systems, which are presented in Section 5. Our article concludes in Section 6 with a discussion about open problems in models for head-to-head outcomes and rating systems.

2. MODELS FOR HEAD-TO-HEAD COMPETITION

The foundation of most probabilistic modeling and development of rating systems for head-to-head competition is the linear paired comparison model (David 1988, Cattelan 2012). Consider a set of n competitors who are to engage in competition, and suppose that competitor $i = 1, \ldots, n$ has a strength parameter θ_i. Let Y_{ij} be a binary outcome of a game played between i and j with

$$Y_{ij} = \begin{cases} 0 \text{ if } i \text{ loses to } j \\ 1 \text{ if } i \text{ defeats } j. \end{cases} \qquad 1.$$

The linear paired comparison model assumes that

$$\Pr(Y_{ij} = 1) = F(\theta_i - \theta_j), \qquad 2.$$

where F is a specified monotonically increasing function, mapping to values between 0 and 1, with the constraint that $F(x) = 1 - F(-x)$ for all $x \in \mathbb{R}$. An appeal of the use of linear paired comparison models is that each competitor has its own scalar parameter as the sole determinant of playing strength. Thus, instead of a model with $\binom{n}{2}$ distinct parameters corresponding to each player pair, linear paired comparison models assume only n parameters.

2.1. Bradley–Terry, Thurstone–Mosteller, and Extensions

The most commonly used choices of the function F are the cumulative distribution function for logistic and for normal distributions, commonly known as the Bradley–Terry (Bradley & Terry 1952) model and the Thurstone–Mosteller (Mosteller 1951) model, respectively. Despite the popularization of these models in the early 1950s, these two models were first introduced in the late 1920s by Zermelo (1928) and Thurstone (1927), respectively. The Bradley–Terry model assumes

$$\Pr(Y_{ij} = 1) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} = \frac{1}{1 + \exp(-(\theta_i - \theta_j))}, \qquad 3.$$

and the Thurstone–Mosteller model assumes

$$\Pr(Y_{ij} = 1) = \Phi(\theta_i - \theta_j), \qquad 4.$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Inference for these models from paired comparisons in games and sports can be accomplished using likelihood methods. The models in Equations 3 and 4, as written, are unidentified without constraints; adding a constant to all θ_i, $i = 1, \dots, n$, results in the same probability expression. It is common to assume a linear constraint on the θ_i to resolve the nonidentifiability. This is typically done either by assuming $\sum_{i=1}^{n} \theta_i = 0$ or by setting $\theta_k = 0$ for some k. Alternatively, regularizing the log-likelihood, e.g., using a ridge or lasso penalty, or by fitting the models in a Bayesian setting with shrinkage priors, can also resolve the nonidentifiability.

The specification in Equation 2 can be generalized so that the θ_i can be a function of covariates. For example, letting $\boldsymbol{x}_i$ be a vector of endogenous covariates to competitor i (i.e., factors that are intrinsic to competitor i's ability), $\boldsymbol{\beta}$ be a corresponding set of covariate effects, and β_{0i} be an intercept specific to competitor i, we can let $\theta_i = \beta_{0i} + \boldsymbol{x}_i'\boldsymbol{\beta}$. In this case, the model in Equation 2 becomes

$$\Pr(Y_{ij} = 1) = F(\theta_i - \theta_j) = F(\beta_{0i} - \beta_{0j} + (\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta}). \qquad 5.$$

The β_{0i} in this model are unidentified, so strategies using a linear constraint or regularization, as with Equations 3 and 4, would need to be employed. In this model, the summary of the ability for competitor i is not β_{0i}, but rather $\beta_{0i} + \boldsymbol{x}_i'\boldsymbol{\beta}$, because the covariates are intrinsic to the competitor's strength.

Two extensions to these linear paired comparison models are worthy of mention. First, in many paired comparison settings, and in games and sports in particular, a decisive result may not be the only possible outcome; it is common for a paired comparison to result in a tie. Ties are common in games like chess and occur with some regularity in soccer, cricket, and, until recent rule changes, hockey. The most commonly used extensions to the Bradley–Terry model that explicitly model the probability of a tie are due to Davidson (1970). The model introduces a parameter that attenuates the frequency of ties relative to the competitor strengths. For a fixed value of the parameter, the probability of a tie is maximized when the competitor strength parameters are equal and decreases as the difference between the competitors' strengths increases. A characterizing feature of this model over other models for ties is that the ratio of the probability of winning to the probability of losing does not depend on the tie parameter. Thus, conditional on the game being decisive, the probability of a win exactly follows the Bradley–Terry model as specified in Equation 3.

A second extension to the Bradley–Terry model that acknowledges a tie as a possible outcome was developed by Rao & Kupper (1967). Their model, like the Davidson (1970) model, also introduces a parameter that governs the frequency of ties relative to the competitor strengths. In the Rao & Kupper (1967) model, the tie parameter can be understood in the following way, relying on an alternative representation of the Bradley–Terry model. Suppose W_{ij}, a latent measure of i outperforming j, is generated from a logistic distribution centered at $\theta_i - \theta_j$. In this representation, i wins if $W_{ij} > 0$, and j wins otherwise. Then $\Pr(W_{ij} > 0) = 1/(1 + \exp(-(\theta_i - \theta_j)))$ is the Bradley–Terry probability in Equation 3. The modification for ties in the Rao & Kupper (1967) model assumes that if W_{ij} is sufficiently close to 0, then the game results in a tie. More specifically, with tie parameter ν, the probability of a tie is determined from $\Pr(-\nu < W_{ij} < \nu)$. The probability of a win and a loss are computed from $\Pr(W_{ij} \geq \nu)$ and $\Pr(W_{ij} \leq -\nu)$, respectively. Larger values of ν correspond to greater probabilities of a tie. A similar threshold-based extension for ties in the context of the Thurstone–Mosteller model was developed by Glenn & David (1960).

In addition to ties, another common extension to linear paired comparison models is the incorporation of an order effect (i.e., the effect for one object of a pair being presented first) or, in the context of sports, a home-field advantage. The most frequently used extension of linear paired comparison models to account for a home-field advantage is to add a parameter δ to the difference in ability parameters $\theta_i - \theta_j$ that reflects the advantage for competing on the home field of i. Thus, the linear paired comparison model becomes

$$\Pr(Y_{ij} = 1) = F(\theta_i - \theta_j + \delta) \qquad 6.$$

when the game is played on the home field of *i*. This approach was developed in the context of the Bradley–Terry model by Davidson & Beaver (1977), and then by Sadasivan (1983) in the context of the Thurstone–Mosteller model. Because the extension of the Bradley–Terry model due to Davidson (1970) involves probabilities of game outcomes as a function of the strength parameters θ_i and θ_j that can be reexpressed only as a function of $\theta_i - \theta_j$, it is straightforward to extend this tie model to include order effects by replacing $\theta_i - \theta_j$ with $\theta_i - \theta_j + \delta$. Indeed, this is precisely the order effect extension also discussed and developed by David (1988).

Direct applications of these models and their extensions have been used in measuring ability in games and sports over the years. Koehler & Ridpath (1982) fit a Bradley–Terry model to NBA game outcomes to measure team strengths, as well as applying the extension by Davidson & Beaver (1977) to measure the extent of a home-court advantage. An application to tennis, modeling individual games within a set via the Bradley–Terry model, was investigated by McHale & Morton (2011). In their approach, the likelihood contributions of game outcomes were exponentially downweighted as a function of time elapsed. Tsokos et al. (2019) investigated models for soccer game outcomes, including the tie extension by Davidson (1970) to the Bradley–Terry model. Similarly, Whelan & Klein (2021) applied the model of Davidson (1970) to the analysis of college hockey results. Other recent examples include modeling test match cricket team strength using the Bradley–Terry model with a home-field advantage parameter (Dewart & Gillard 2019), and a Bayesian adaptation of the Bradley–Terry model to model Major League Baseball team strengths (Phelan & Whelan 2018).

2.2. Application to National Basketball Association Game Outcomes in the 2017–2018 Season

We demonstrate the application of the Bradley–Terry and Thurstone–Mosteller models to evaluating team strength during the 2017–2018 NBA regular season. We fit each model under the zero-sum constraint $\sum_{i=1}^{n} \theta_i = 0$, and estimate the 30 team parameters via maximum likelihood. These models do not include home-field advantage parameters. This amounts to fitting a constrained logistic regression for the Bradley–Terry model, and a constrained probit regression for the Thurstone–Mosteller model. The results of fitting the model on all 1,230 games played among 30 teams is shown in **Table 1**, sorted according to the Bradley–Terry estimates.

Several features of the results are worthy of comment. The team rankings according to the Bradley–Terry and Thurstone–Mosteller models are nearly identical, except for the Orlando Magic and the Dallas Mavericks, whose order is reversed in the Thurstone–Mosteller model (the difference in the strength parameter estimates in each model is nearly indistinguishable). Similarly, the rank order of the parameter estimates for both the Bradley–Terry and Thurstone–Mosteller models is consistent with the number of won games out of 82 by each team. For both models, the Portland Trail Blazers had higher estimated strength parameters than the Cleveland Cavaliers, but the latter team won one more game. This is likely the result of the Trail Blazers having a more challenging schedule with tougher opponents, on average. A similar phenomenon happened between the Magic and the Mavericks under the Thurstone–Mosteller model.

The range of strength estimates for the two models indicates the degree to which teams are favored to win. According to the Bradley–Terry model, the probability that the top team (the Houston Rockets) would defeat the worst team (the Phoenix Suns) is approximately $1/(1 + \exp(-(1.392 - (-1.089)))) = 0.923$. For the Thurstone–Mosteller model, the probability

Table 1 Bradley–Terry and Thurstone–Mosteller estimates and standard errors, and total number of won games for the 2017 National Basketball Association regular season

Team	Bradley–Terry estimate (standard error)	Thurstone–Mosteller estimate (standard error)	Total wins out of 82
Houston Rockets	1.392 (0.271)	0.831 (0.156)	65
Toronto Raptors	0.956 (0.247)	0.593 (0.147)	59
Golden State Warriors	0.928 (0.244)	0.568 (0.146)	58
Boston Celtics	0.719 (0.238)	0.441 (0.143)	55
Philadelphia 76ers	0.578 (0.232)	0.351 (0.141)	52
Portland Trail Blazers	0.440 (0.229)	0.266 (0.140)	49
Cleveland Cavaliers	0.428 (0.230)	0.261 (0.140)	50
Utah Jazz	0.423 (0.228)	0.254 (0.139)	48
New Orleans Pelicans	0.397 (0.229)	0.237 (0.139)	48
Oklahoma City Thunder	0.378 (0.228)	0.230 (0.139)	48
Indiana Pacers	0.350 (0.228)	0.216 (0.139)	48
Minnesota Timberwolves	0.341 (0.228)	0.198 (0.139)	47
San Antonio Spurs	0.330 (0.228)	0.207 (0.139)	47
Denver Nuggets	0.274 (0.227)	0.163 (0.139)	46
Milwaukee Bucks	0.140 (0.225)	0.089 (0.138)	44
Miami Heat	0.106 (0.225)	0.063 (0.138)	44
Washington Wizards	0.104 (0.225)	0.061 (0.138)	43
Los Angeles Clippers	0.075 (0.226)	0.053 (0.138)	42
Detroit Pistons	−0.108 (0.225)	−0.066 (0.138)	39
Charlotte Hornets	−0.281 (0.227)	−0.176 (0.139)	36
Los Angeles Lakers	−0.307 (0.228)	−0.189 (0.139)	35
New York Knicks	−0.642 (0.234)	−0.392 (0.142)	29
Brooklyn Nets	−0.696 (0.236)	−0.432 (0.143)	28
Sacramento Kings	−0.733 (0.238)	−0.439 (0.143)	27
Chicago Bulls	−0.748 (0.238)	−0.466 (0.143)	27
Orlando Magic	−0.896 (0.242)	−0.547 (0.145)	25
Dallas Mavericks	−0.898 (0.246)	−0.541 (0.146)	24
Atlanta Hawks	−0.927 (0.245)	−0.561 (0.146)	24
Memphis Grizzlies	−1.034 (0.252)	−0.603 (0.148)	22
Phoenix Suns	−1.089 (0.255)	−0.673 (0.150)	21

is estimated as $\Phi(0.831 - (-0.673)) = 0.934$. This large probability suggests that NBA teams in 2017 (and similarly in other years) have a substantial range of strengths.

The standard errors of the estimates indicate a fair amount of uncertainty about the actual team strengths. For example, with the typical Bradley–Terry strength parameter standard error of about 0.23, a 95% confidence interval has a margin of error of roughly 0.45. This indicates that large sets of teams, especially those in the middle of the ranking, are nearly indistinguishable in strength, when accounting for the uncertainty in the estimates. However, it is worth noting that covariances of the strength estimates tend to be positive (and of greater magnitude for teams who compete), so that differences in strength tend to be more reliably estimated than individual strength parameters.

We also fit a Bradley–Terry and Thurstone–Mosteller model that includes a home-field advantage (HFA) parameter (results not shown). The order of teams for the Bradley–Terry model with the HFA parameter is identical to that of the version without an HFA parameter. For the Thurstone–Mosteller model with an HFA parameter, a few pairs of teams swap their order relative to the results of the version without an HFA parameter, though accounting for the estimation uncertainty, the swaps are not practically meaningful. The HFA parameters for the two models are significantly positive [0.384(0.064) for the Bradley–Terry model, and 0.230(0.038) for the Thurstone–Mosteller model], validating the well-known phenomenon that teams experience an advantage when playing on their home court. For two teams that are evenly matched on a neutral court, the Bradley–Terry and Thurstone–Mosteller models estimate that the probability of a team winning on their home court is $1/(1 + \exp(-0.384)) = 0.595$ and $\Phi(0.230) = 0.591$, respectively.

To assess which of the Bradley–Terry and Thurstone–Mosteller models are a better fit to the data, a cross-validation analysis can be performed that evaluates the predictability of each model on withheld data. One way to accomplish this task is to perform leave-one-out cross-validation a total of 1,230 times, fitting each model on 1,229 game outcomes and evaluating the predicted probability of the left-out game. Letting p_i^* be the predicted probability of the home team winning game i, which was left out of the modeling, and y_i be the binary outcome for game i relative to the home team, we can compute the average log-loss (Good 1952, Gneiting & Raftery 2007) as

$$\text{log-loss} = \frac{-1}{1,230} \sum_{i=1}^{1,230} (y_i \log\ p_i^* + (1 - y_i) \log(1 - p_i^*)). \qquad 7.$$

When applied to the Bradley–Terry and Thurstone–Mosteller models that include an HFA parameter, the log-loss for each model is evaluated to be 0.2876337 for the Bradley–Terry model, and 0.2880857 for the Thurstone–Mosteller model. Thus, the Bradley–Terry model, with a lower log-loss, is a slightly better fit.

3. TIME-VARYING PAIRED COMPARISON MODELS

An important extension to linear paired comparison models is the assumption that the strength parameters may be time-varying. It is difficult to imagine sports settings in which game outcomes are collected over a span of time, but in which competitor strengths do not correspondingly evolve. Most serious analyses of sports outcome data, with an aim toward measuring competitor strength or making game outcome forecasts, involve some acknowledgment that player or team strengths are changing over time.

Two overarching approaches to modeling time-varying strengths have been proposed. The first involves modeling competitor strength over time through a nonstochastic function in conjunction with the paired comparison model for game outcomes. The second is to assume that a competitor's strength evolves through a stochastic process. In both cases, we assume that competitor i at time t has strength parameter θ_{it}, and that

$$\Pr(Y_{ijt} = 1) = F(\theta_{it} - \theta_{jt}), \qquad 8.$$

where, again, F is a specified continuous cumulative distribution function with real support.

3.1. Nonstochastic Extensions

The nonstochastic approach assumes the existence of a smooth function f_i for competitor i for which

$$\theta_{it} = f_i(t|\gamma), \qquad 9.$$

where γ is a parameter vector common to all competitors. In principle, likelihood-based methods can be employed by combining Equation 8 with Equation 9 to obtain the full likelihood specification, and performing inference for the f_i and γ through standard approaches.

Many choices exist for the functional form of f_i. While we are not aware of previous work that has assumed the f_i to be as simple as polynomial functions, Araki et al. (2019) have developed a time-varying Bradley–Terry approach using piecewise polynomial splines. Nonparametric splines, including smoothing splines, have been suggested, e.g., in Baker & McHale (2014), though have rarely been used. As one example, Bong et al. (2020) have assumed the f_i are kernel smoothers in extending the Bradley–Terry model. Probably the most common choice of the f_i is to assume that they follow barycentric rationale interpolants (BRIs) (Taylor 1945, Berrut et al. 2011). This approach has been popularized in several papers by Baker and McHale over the past ten years (Baker & McHale 2014, 2015, 2017), and has also been adopted by Krese & Štrumbelj (2021). This approach, which Baker & McHale (2014) argue is an appealing choice given the simplicity of the expressions compared with other flexible functions, assumes

$$\theta_{it} = \frac{\sum_{k=1}^{n_i} w_{ik}\hat{\lambda}_{ik}(t - t_{ik})}{\sum_{k=1}^{n_i} w_{ik}/(t - t_{ik})}, \qquad 10.$$

where $\hat{\lambda}_{ik}$ is an estimated value of θ_{it} at $t = t_{ik}$. An algorithm to determine the weights w_{ik} and location of the knots k is described by Floater & Hormann (2007).

Nonstochastic extensions to linear paired comparison models have been applied to a variety of sports settings. Baker & McHale (2014, 2017) developed an approach using BRIs to measure time-varying strength of tennis players. Bong et al. (2020) applied their kernel smoothing approach to the analysis of NFL football data over time. Finally, Krese & Štrumbelj (2021) demonstrated the development of the Bradley–Terry model with BRIs to measure player strength in sumo wrestling.

3.2. Stochastic Extensions

A complementary approach to extending linear paired comparison models through smooth functions over time is the specification of a stochastic process on the strength parameters. This approach is more generally known as state-space modeling (Durbin & Koopman 2012), which treats the head-to-head outcome model conditional only on parameters fixed in time, and a separate (usually Markovian) model that governs the evolution of the ability parameters from one time period to the next. Given that linear paired comparison models assume that the strength parameters θ_{it} are real-valued, a common choice is to assume first-order autoregressive model on the θ_{it}, that is,

$$\theta_{i,t+1} = \rho\theta_{it} + \varepsilon_{i,t+1}, \qquad 11.$$

where $|\rho| < 1$ is an autoregressive parameter, and $\varepsilon_{i,t+1} \sim \mathrm{N}(0,\sigma^2)$ with unknown innovation variance σ^2. This component to the model recognizes that the strength parameters may undergo random innovations over time, as opposed to a smooth functional trajectory as assumed by the conventional nonstochastic approaches.

Likelihood-based inference can be performed noting that

$$\mathrm{L}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T, \rho|\boldsymbol{y}) \propto \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{\theta}_t) \prod_{t=2}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \rho), \qquad 12.$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{y}_t$ are the vectors of ability parameters and game outcomes, respectively, during time t; $p(\boldsymbol{y}_t|\boldsymbol{\theta}_t)$ is the joint mass function for all game outcomes during time t; and $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \rho)$ is the joint density of innovations through a Markov transition.

One way to interpret the state-space modeling approach is to consider two extremes: a setting in which a model treats a team's ability as constant across time, and a setting in which a model acts as though no connection exists between team strength parameters over time. In the former case, the analyst would likely pool together all the data and fit a model, like the Bradley–Terry or Thurstone–Mosteller model, to all the game outcomes, ignoring any possible time variation in team abilities. In the latter case, the analyst would likely analyze the games played during each time period separately and summarize strength estimates based on these separate analyses. The problem with the pooling approach is that it is likely biased, resulting in single strength estimates for each team that are constant over time, even if the team's ability is changing. The approach of analyzing game data within each time period separately suffers from potentially high variance—that is, the model uses only game outcome data within the time period to estimate ability, but does not take advantage of the likely consistency in a team's ability over time. The state-space approach can be viewed as a compromise between these two extremes. The model in Equation 11 essentially assumes that two strength parameters consecutive in time, θ_{it} and $\theta_{i,t+1}$, are distinct parameters, but that they cannot be too far apart. The innovation variance, σ^2, is the main parameter that characterizes the magnitude of the difference between strength parameters over time. When $\sigma^2 = 0$ (and $\rho = 1$), the state-space model becomes the approach where a team's strength is constant over time. When $\sigma^2 \to \infty$, then the model acts as though the strength parameters have no connection over time. In usual application, σ^2 will be inferred to be a positive (finite) parameter, and therefore a compromise between the two extremes.

The approach to inference for the state-space approach is almost always performed within a Bayesian framework. This means that a prior distribution on the θ_{i1} for $i = 1, \dots, n$ needs to be assumed for a full model specification. Without additional information, assuming independent prior distribution components $\theta_{i1} \sim \mathrm{N}(0, \sigma_1^2)$, where σ_1 would be chosen to be large to reflect initial uncertainty in a team's ability, is a sensible choice. Alternatively, with prior information about teams' abilities, an informative prior distribution may be assumed for the θ_{i1}.

One benefit to incorporating the state-space approach in a Bayesian setting is that if the goal is to obtain inferences for the θ_{iT}, the most current set of team strengths, on an ongoing basis, it is unnecessary to perform a full Bayesian analysis every time new data are recorded. Instead, the posterior distribution can be sequentially updated via a recursive updating algorithm. Let D_t denote all game outcomes through time period t, and assume σ^2, the innovation variance, is either known or estimated in advance. Suppressing the conditioning on σ^2, now suppose we have a prior distribution on strength parameters for time period t, $p(\boldsymbol{\theta}_t | D_{t-1})$. After observing game outcomes, $\boldsymbol{y}_t$, at time t, the prior can be updated to the posterior distribution via Bayes' rule as $p(\boldsymbol{\theta}_t | D_t)$. Then, to obtain the prior distribution for strength parameters at the next time period, a standard Bayesian calculation can be performed by determining $p(\boldsymbol{\theta}_{t+1} | D_t) = \int p(\boldsymbol{\theta}_t | D_t) p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \sigma^2) \mathrm{d}\boldsymbol{\theta}_t$, where $p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \sigma^2)$ is the density of the innovation distribution. In practice, if the prior distribution for time t is assumed normal, the posterior distribution may be approximated by a normal distribution as in Glickman (1999), noting that the prior is not conjugate with the likelihood. Such a sequential updating procedure is far more efficient than reanalyzing the entire posterior when new game data are observed.

Linear paired comparison state-space models can be viewed as instances of dynamic generalized linear models (West et al. 1985). Early examinations of these paired comparison models were carried out by Glickman (1993), who developed a Markov chain Monte Carlo (MCMC) algorithm for posterior inference and demonstrated the approach to the analysis of chess player strength, and Fahrmeir & Tutz (1994), who developed an efficient quasi-Newton algorithm for posterior mode estimation applied to soccer game outcomes, assuming an ordered logit model for game outcomes assuming ties. More recently, Knorr-Held (2000) assumed the same basic model as that developed by Fahrmeir & Tutz (1994) but used an MCMC approach to model fitting. This approach was also applied to the analysis of soccer game outcomes. Cattelan et al. (2013) developed a state-space linear paired comparison model applied to the analysis of NBA basketball games. Finally, Gorgi et al. (2019) demonstrated the application of state-space modeling to the analysis of tennis game outcomes.

4. RATING SYSTEMS

The approaches to paired comparison models with time-varying parameters described in Sections 3.1 and 3.2, and their associated analyses, are appropriate in scenarios where the number of competitors and the number of time periods is not too large. This is because the likelihood-based analyses depend only on a limited number of parameters in such situations, so standard optimization methods or Monte Carlo sampling from the posterior distribution is a tractable problem. However, many situations, such as league play with many teams or gaming organizations (especially online gaming) with thousands of competing participants, pose challenges for standard numerical approaches. Additionally, it is often of interest in league play or in online gaming for inference to focus only on current playing strength, and of far less interest to estimate abilities in the past, a by-product of the likelihood-based approaches. The advent of rating systems as methods to measure current strength from head-to-head competition has become pervasive in sports and gaming applications, and it has mostly supplanted the need for likelihood-based approaches in settings with large numbers of competitors.

The main distinction between rating systems and full likelihood-based methods is that the former is primarily focused on local updates in the strength parameters. Rather than attempting to make inferences on all strength parameters across all competitors over all time periods in one analysis, rating systems implement a simpler procedure in which strength estimates at the start of a time period t are treated as known, and then, based on game results during period t, the strength estimates are updated and typically used at the start of period $t + 1$. The associated computations are typically referred to as filtering, and the process itself is recursive in its application. Because these systems update each competitor's rating in parallel with all other competitors, the computation scales linearly with the number of competitors. Invariably, the computational tractability in this type of filtering approach is far better than approaches based on using the full likelihood. However, this usually comes at the expense of accuracy and faithfulness to the underlying probability model.

4.1. The Elo Rating System

Arguably the first serious rating system to probabilistically measure strength in head-to-head competition is the approach developed by Elo (1978). Elo developed his rating system in the early 1950s, roughly contemporaneously with the development of the Bradley–Terry model. Despite the strong connections between the Bradley–Terry model and Elo's system, it is unclear whether Elo was aware of Bradley and Terry's work.

The Elo rating system assumes, at its core, the Bradley–Terry probability model, albeit with a linear transformation of the strength parameters relative to the Bradley–Terry model. Suppose $\hat{\theta}_{it}$ is an estimate of the strength parameter for competitor i during time period t, following the Bradley–Terry model as specified in Equation 3. For each $i = 1, \ldots, n$, let

$$R_{it} = 1500 + \left(\frac{400}{\log 10}\right)\hat{\theta}_{it} = 1500 + 173.72\ \hat{\theta}_{it} \qquad 13.$$

be the Elo rating of player i at time t, and define

$$\mathrm{We}_{ijt} = \frac{1}{1 + 10^{-(R_{it} - R_{jt})/400}} \qquad 14.$$

be the "winning expectancy" of competitor i over competitor j at time t. For games with binary outcomes, the Elo winning expectancy is the estimated Bradley–Terry model probability of *i* defeating *j* at time *t*. For games with ties, the expression in Equation 14 is treated as an estimated expected score.

The Elo updating algorithm updates ratings at time t to $t+1$ based on game outcomes during time period t. The algorithm assumes known ratings $R_{1t}, \ldots, R_{nt}$ at the start of time period t. For competitor i who competed during time period t, suppose that their M opponents are $j_1, \ldots, j_M$. The rating for competitor i is updated using the formula

$$R_{i,t+1} = R_{it} + K \sum_{m=1}^{M} \left(y_{ij_mt} - \mathrm{We}_{ij_mt} \right), \qquad 15.$$

where K is a constant that reflects the degree to which game results impact the change in rating. The formula in Equation 15 is applied in parallel for all competitors who compete during time period t. While Equation 15 resembles a possible linear approximation to a Bayesian update of a prior rating to a posterior rating, the Elo system was not derived in connection to a Bayesian analysis, and can be viewed only as a degenerate special case of Bayesian updating, as mentioned in Section 4.2. The Elo system also does not provide uncertainty measures of ratings. A derivation of the formula due to Elo is provided in the **Supplemental Appendix**.

The Elo updating formula in Equation 15 applies typically after competitor i has already completed some specified number of games. Prior to that number, and most obviously when a player first competes and does not possess a rating, a different set of formulas is used to determine a player's rating. Suppose that a player without a rating competes against M players $j_1, \ldots, j_M$, and assume the opponents' ratings are $R_{j_1}, \ldots, R_{j_M}$. To determine a provisional rating based on results against these opponents, one method that has been suggested (Elo 1978) is to compute

$$R_i = \frac{\sum_{m=1}^{M} (R_{j_m} + 400(2y_{ij_m} - 1))}{M}, \qquad 16.$$

which is the average of the opponents' ratings plus or minus 400, adding 400 for opponents who defeated the player, and subtracting 400 from the rating of opponents who the player defeated. If an opponent did not possess a rating, then that rating would be omitted from the computation in Equation 16.

The provisional rating formula has a notable undesirable feature, namely that it can produce ratings that decrease after defeating an opponent. For example, suppose a player defeats, draws, and loses to players rated 1400, 1500, and 1600, respectively. The resulting provisional rating would be the average of $(1400 + 400)$, $(1500 + 0)$, and $(1600 - 400)$, or 1500. Now suppose the player defeats a fourth opponent rated 1000. Based on the four games, the player's rating would be the average of $(1400 + 400)$, $(1500 + 0)$, $(1600 - 400)$, and $(1000 + 400)$, or 1475. Thus, with the inclusion of a win against a low-rated opponent, we have the result that the player has a lower rating. The version of the Elo system implemented by US Chess (Glickman & Doan 2024) resolved this difficulty by recognizing that the formula in Equation 16 is closely connected to the maximum likelihood solution to a model that assumes the probability of competitor i defeating j_m is given by

$$\Pr(y_{ij_m} = 1 | R_i, R_{j_m}) = \begin{cases} 0 \text{ if } R_i \le R_{j_m} - 400 \\ 0.5 + (R_i - R_{j_m})/800 \text{ if } R_{j_m} - 400 < R_i < R_{j_m} + 400 \\ 1 \text{ if } R_i \ge R_{j_m} + 400. \end{cases} \qquad 17.$$

Using Expression 17 for the probability of a win, the likelihood for R_i can be constructed as products of probabilities of the observed game outcomes. The contribution of a tie can be included into the likelihood as $\sqrt{\Pr(y_{ij_m}=1|R_i,R_{j_m})(1-\Pr(y_{ij_m}=1|R_i,R_{j_m}))}$. The optimizing value of R_i can be determined numerically, most simply by a bisection algorithm. Because the likelihood may be maximized for ranges of values of R_i (for example, defeating an opponent rated 1000 and losing to an opponent rated 2000 will result in values of $1400 \leq R_i \leq 1600$, all of which maximize the likelihood) rather than a single value, a prior estimate of R_i is usually assumed, and the value of the MLE closest to the prior estimate is chosen as the final point estimate.

Another simple approach to rating new competitors in the Elo system is to rely entirely on the Elo updating formula in Equation 15. In this approach, all players would start with the same rating, typically an average rating like 1500, though with extra information, other initial ratings may be chosen on a case-by-case scenario. Instead of using the K factor normally used, a larger K factor could be used instead for the first time period. The use of a larger K factor reflects the understanding that a competitor's rating needs to respond to game results in the first time period more so than in subsequent time periods. In the second and following time periods, the K factor reverts to a lower value. We apply this approach in our data analyses in Section 4.3.

The Elo rating system, and variations of the system, has been used and implemented in many settings. The original version of the Elo system was adopted by US Chess in 1960, and even though the current US Chess system is much more complicated than the original system, the basic updating algorithm follows Elo's formulation. The International Chess Federation (FIDE) has also used versions of the Elo system since 1970. Numerous gaming and sports leagues have also adopted versions of the Elo system, including organized backgammon (Keith 2019), the Fédération Internationale de Football Association (FIFA) Men's and FIFA Women's world rankings for soccer teams (FIFA 2023a,b), the online game League of Legends prior to Season 3 (League of Legends Wiki 2022), the online game Age of Empires (Age of Empires DE Team 2021), online poker (GGPoker 2023), and USA Pickleball (USA Pickleball 2018), to name a few.

4.2. Glicko and Glicko-2 Rating Systems

Glicko and Glicko-2 are a popular pair of rating systems that were developed in the 1990s by Dr. Mark Glickman that both approximate likelihood-based inference from Bayesian dynamic models of the type described in Section 3.2. The Glicko and Glicko-2 systems were derived as linear approximations to Bayesian sequential updating from the prior at time period t to the posterior at time period t, and then to the prior at time period $t + 1$. Descriptions of these algorithms are available in documents on Dr. Glickman's ratings page (Glickman 2022). The technical development of the Glicko and Glicko-2 systems can be found in Glickman (1999) and Glickman (2001), respectively. The main difference between the Glicko and Glicko-2 systems is in the assumed stochastic process by which player strengths evolve over time. In the Glicko system, strength parameters evolve through a Gaussian process, whereas in the Glicko-2 system, the strength parameters evolve through a Gaussian-based stochastic volatility process. Both systems can be viewed as modifications of the Elo updating algorithm that specifically account for the uncertainty in playing strength estimation.

The Glicko and Glicko-2 systems involve recursive updates, like the Elo system. The strength parameter for player i at time t, θ_{it} (for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$), is assumed to have a normal prior distribution,

$$\theta_{it} \sim \mathrm{N}(\mu_{it}, \tau_{it}^2). \qquad 18.$$

Once games are played during time period t, an approximate normal prior distribution for time period t+1 can be determined by an analysis that acknowledges the game results as well as the innovation in strengths from time t to t+1. Specifically, the Glicko system assumes an innovation component of the form in Equation 11 with $\rho = 1$ and with variance parameter σ^2. The Glicko-2 system assumes that the innovation variance, now indexed by player i and time t, σ_{it}^2, undergoes a discrete-time log-normal process. In the Glicko system, the approximation formulas are in closed form. For the Glicko-2 system, iterative computation is necessary to solve one-dimensional optimizations for each player whose parameter update is being computed. Glickman (1999) illustrates that the Elo system is, in fact, a special case of the Glicko system when the normal prior variances of player strengths are 0, implying that the strengths are known with certainty.

Incorporating uncertainty into a rating algorithm has important implications for measuring playing strength that are absent in the Elo system. First, because the Glicko and Glicko-2 systems produce approximate normal posterior distributions for competitor strength, credible intervals for strength can be determined in a straightforward manner that reflects the uncertainty in estimation, and can be used to compute posterior predictive distributions for future game outcomes. Furthermore, unlike the Elo system, both the Glicko and Glicko-2 systems do not require special alternative computations for unrated or provisionally rated players. As long as a prior distribution before a rating period can be specified, the algorithm does not need to distinguish between provisional and established players. For competitors new to a system, assuming a diffuse normal prior centered at a mean that may depend on relevant information (e.g., in chess, based on a player's age) would be an appropriate choice. Second, players who compete infrequently or have played very few games would likely have much less reliable strength estimates than players who compete regularly and have been playing for years. The Elo system does not distinguish between these two types of players. In the former situation, the Glicko and Glicko-2 systems typically involve appreciable changes to the mean strength given that the player's strength is uncertain, but in the latter situation the Glicko and Glicko-2 systems tends to involve only small changes to the mean strength.

There is an important difference between assuming a Gaussian process for strength parameters in the Glicko system and the stochastic volatility process for strength parameters in the Glicko-2 system. Because the Glicko system allows only normal innovations to the strength parameters, it may be unable to capture sudden bursts in improvement, especially with competitors who compete regularly and whose variance of their strength distribution is small. With the stochastic volatility process in the Glicko-2 system, inconsistencies in a player's game results with their prior distribution result in an innovation to the prior variance. This has the effect of allowing the mean strength parameter to change substantially, despite having a small prior variance. In essence, the Glicko-2 system acknowledges game results that are inconsistent with one's prior distribution by increasing the strength distribution variance as well as changing the mean substantially. The end result is an approximating normal prior distribution for the next time period that has appreciably changed in mean, but reflects the new uncertainty of this mean with a large prior variance.

The first organization to adopt the Glicko system was the Free Internet Chess Server (FICS) in 1995 (FICS 2008). Other online chess organizations that have adopted these systems include Chess.com (Allebest 2018), which has implemented the Glicko system, and Lichess.org (2023), which has implemented a variant of the Glicko-2 system. The Australian Chess Federation for over-the-board chess has also adopted a variant of the Glicko-2 system (ACF 2013). In addition to chess, other gaming systems have adopted the Glicko and Glicko-2 systems. These include Counter-Strike: Global Offensive, which uses the Glicko-2 system (Medado 2021); Defense of the Ancients 2, which uses the Glicko system (Çakır et al. 2024); online Go (Noek 2017); and many others.

4.3. Application to National Basketball Association Game Outcomes

We demonstrate the application of the Elo and Glicko rating systems to regular season NBA game outcomes from the 2004–2005 season to the 2017–2018 season. The procedure we describe below can be analogously applied to the Glicko-2 rating system with similar results to the Glicko system. For the Elo and Glicko analyses, we treat each season as its own time period within which team strengths are assumed to be unchanging, but can evolve over time between seasons. Both analyses involve tuning parameters specific to each system, the K factor in the Elo system, and the innovation variance parameter σ^2 in the Glicko system. Our analyses uses the `PlayerRatings` R package (Stephenson & Sonas 2020) implementation for the Elo and Glicko rating systems. For each set of analyses, we used the 2004–2005 through 2014–2015 seasons as training data and the 2015–2016 through 2017–2018 seasons as validation game outcomes to optimize the tuning parameters as described below.

Several papers have explored the comparison between the Elo and Glicko systems for various applications. Veček et al. (2014) compared the performance of evolutionary algorithms using the Elo and Glicko-2 algorithms, along with several others. They found that Glicko-2 was most appropriate for their application. Ingram (2021) applied the Elo and Glicko systems to professional men's tennis game outcomes and found that Elo slightly outperformed Glicko. In contrast, Yue et al. (2022) found that the Glicko system outperformed the Elo system on women's professional tennis game outcomes. It appears that the level of performance may depend on the data to which the rating systems are applied.

In our implementation of the Elo system, we assumed a K factor specific to the first season (2004–2005), and a separate K factor that would be assumed for the subsequent seasons. Based on the results of optimizing over both K factors, we found that the initial K factor had essentially no impact on the final predictability of Elo ratings after the first few seasons over a wide range of candidate values. We set the initial K factor to 30 and optimized only for the K factor that was assumed for subsequent seasons. The optimization criterion we used was the log-likelihood over game outcomes (which is linearly related to the log-loss) using the Elo winning expectancy formula for all games played in the validation period. As shown in **Figure 1**, the optimal value of K was 6.2.

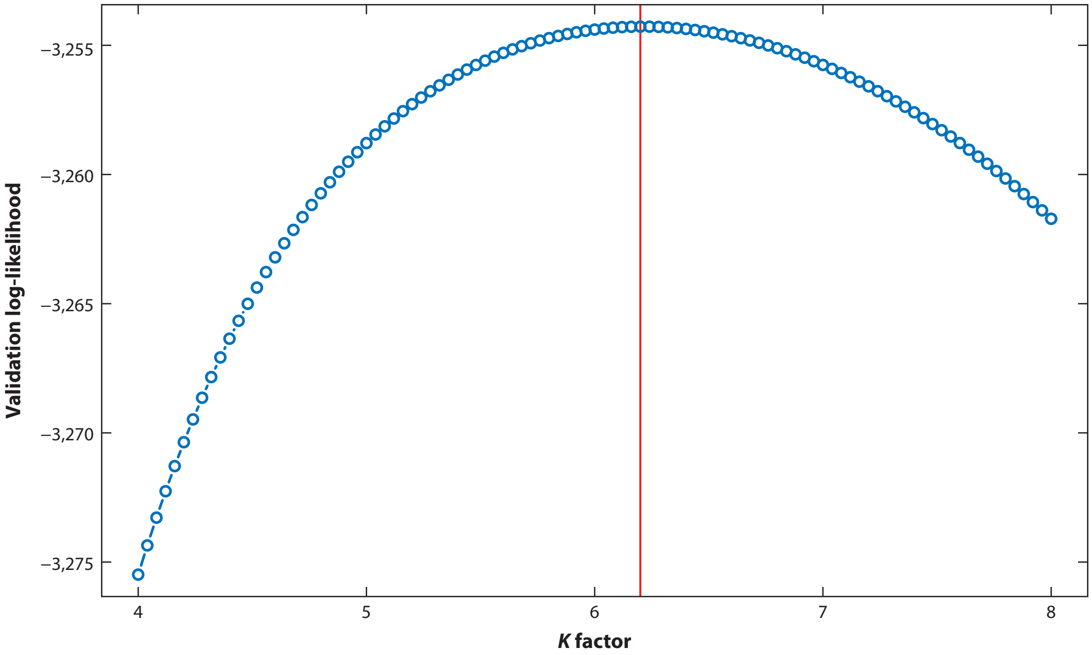

Figure 1

Optimal K factor of 6.2 in the Elo system based on analysis of 2004–2017 National Basketball Association regular season data.

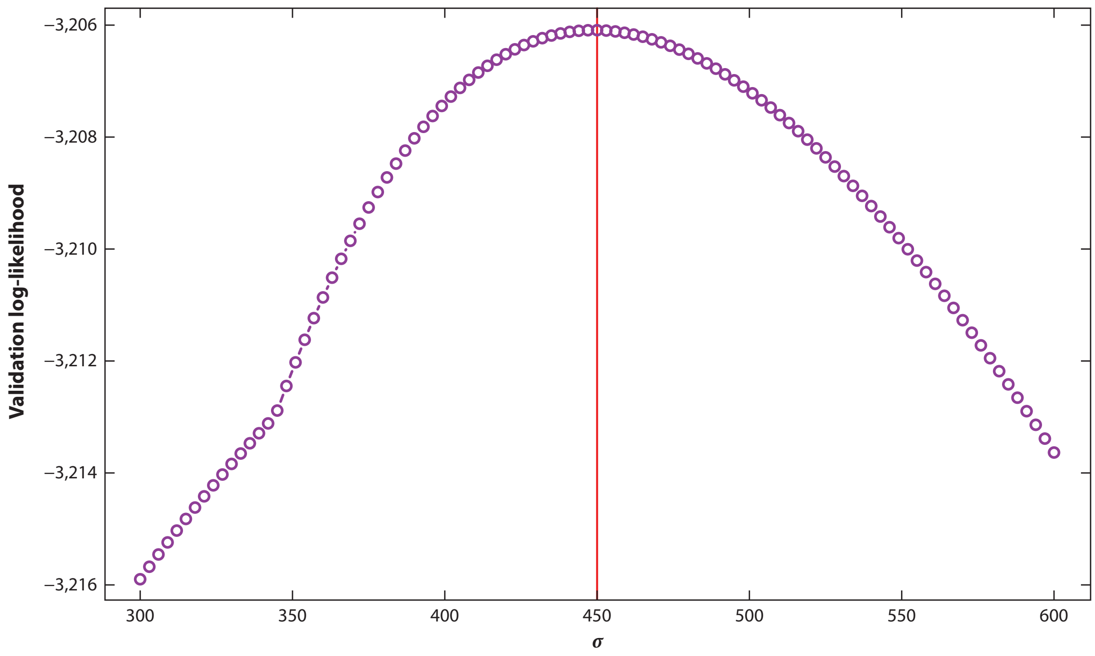

Figure 2

Optimal innovation standard deviation $\sigma = 450$ in the Glicko system based on analysis of 2004–2017 National Basketball Association regular season data.

Similarly, we implemented the Glicko system recognizing that two parameters needed to be optimized: a prior standard deviation common to all teams in 2004 ($\tau_{i1} = \tau_1$) and the innovation standard deviation, σ. Much like the Elo analysis, the results barely depended on the selection of the prior standard deviation for a wide range of values of τ_1, so we set $\tau_1 = 350$, a value recommended by Glickman (2022). We optimized the log-likelihood over game outcomes in the validation set using the expected score of a game, approximately marginalized with respect to the team normal prior distributions. The exact specification of this computation is provided by Glickman (1999, 2022). **Figure 2** shows the log-likelihood over the range of candidate σ values, the optimal being achieved at $\sigma = 450$.

A summary of the team ratings at the end of the 2017 regular season for both optimized Elo and Glicko systems are displayed in **Table 2**. The ratings are generally consistent (with a Pearson correlation of 0.933), but several teams exhibit notable differences between the two rating systems. For example, the Houston Rockets, who had the best record in the 2017–2018 season, have a substantially lower rating in the Elo system compared with the Glicko system. This is also the case for other teams, such as the Toronto Raptors, the Philadelphia 76ers, and the Boston Celtics, but not to the same extent. This is arguably explained by the optimized Elo K factor effectively giving higher weight to previous years' results than the Glicko system does. In fact, the Glicko ratings are nearly linearly related to the Bradley–Terry and Thurstone–Mosteller estimates that are based only on 2017 regular game results, as can be seen in the pairwise scatter plots displayed in **Figure 3**. The correlations among the Glicko, Bradley–Terry, and Thurstone–Mosteller estimates are all above 0.998, while the correlation between the Elo rating and the other three are never higher than 0.947.

Table 2 Elo and Glicko ratings at the end of the 2017 National Basketball Association regular season

Team	Elo rating	Glicko rating
Golden State Warriors	1715	1795
Houston Rockets	1685	1948
Toronto Raptors	1633	1821
San Antonio Spurs	1617	1610
Boston Celtics	1604	1755
Cleveland Cavaliers	1577	1668
Oklahoma City Thunder	1572	1658
Utah Jazz	1569	1662
Portland Trail Blazers	1558	1671
Los Angeles Clippers	1543	1551
Indiana Pacers	1542	1646
New Orleans Pelicans	1533	1674
Denver Nuggets	1529	1629
Washington Wizards	1522	1566
Miami Heat	1513	1574
Minnesota Timberwolves	1509	1652
Milwaukee Bucks	1507	1581
Philadelphia 76ers	1499	1709
Detroit Pistons	1474	1507
Charlotte Hornets	1455	1458
Los Angeles Lakers	1414	1462
Chicago Bulls	1410	1317
Atlanta Hawks	1399	1268
Memphis Grizzlies	1396	1241
Sacramento Kings	1393	1320
Dallas Mavericks	1389	1274
New York Knicks	1388	1344
Orlando Magic	1364	1276
Brooklyn Nets	1354	1356
Phoenix Suns	1336	1217

Teams are sorted in order of their Elo ratings.

The Elo and Glicko rating systems can be used to examine the trajectory of team abilities over time. In **Figures 4** and **5**, we plot the Elo and Glicko ratings, respectively, for the five teams in the NBA Atlantic division (consisting of the Boston Celtics, the Brooklyn Nets, the New York Knicks, the Philadelphia 76ers, and the Toronto Raptors) over the 2004–2005 to 2017–2018 seasons. While the trajectories for individual teams follow similar trends between the two figures, the Elo ratings generally do not change quite as abruptly from season to season. This is consistent with the low optimized *K* factor that prevents changes of large magnitudes across seasons. The change in Glicko ratings tends to be more appreciable for these data, where an increase/decrease of 400–500 rating points between seasons can occasionally occur (e.g., the increase in the Celtics' rating from 2006–2007 to 2007–2008, the latter season being when they won the NBA championship).

The predictive performance of Elo and Glicko can be compared by applying the rating systems optimized on game outcomes through the 2017–2018 season to the games in the 2018–2019

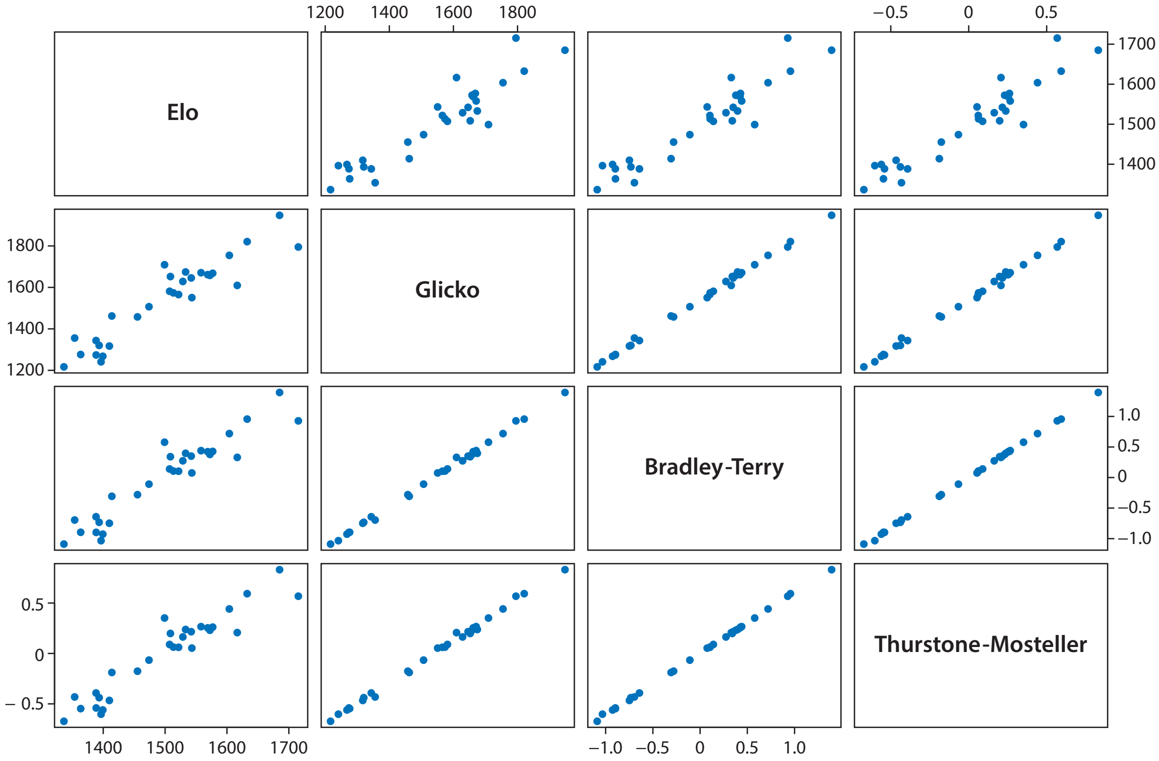

Figure 3

Pairwise plots of 2017 National Basketball Association team strength estimates among Elo, Glicko, Bradley–Terry, and Thurstone–Mosteller models.

season. We computed the (predictive) log-likelihood for the two rating systems over the 1,230 games in the 2018–2019 regular season. For the Elo system, the log-likelihood was −1,841.8, and for the Glicko system, the log-likelihood was −811.8. This large difference suggests that the Glicko system is more predictive of game outcomes than the Elo system for NBA basketball games.

5. MODERN SYSTEMS

Many organizations use versions of the Elo system, such as FIDE, but others have developed more modern systems.

5.1. TrueSkill

Another rating system that has been used in online gaming, particularly for Microsoft products, is TrueSkill. This system was developed in the early 2000s (Herbrich et al. 2006) and later revised to TrueSkill 2 (Minka et al. 2018). In addition to rating players in head-to-head competition, TrueSkill also permits rating individual players within team-versus-team competitions through weighted averages of the player ratings, with the weights proportional to the amount of time a player has played in a game. Methods for measuring player strengths within head-to-head team games have had little attention outside of TrueSkill.

The system assumes the Thurstone–Mosteller model for game outcomes. TrueSkill also acknowledges the possibility of tied games and uses the threshold parameter extension of the

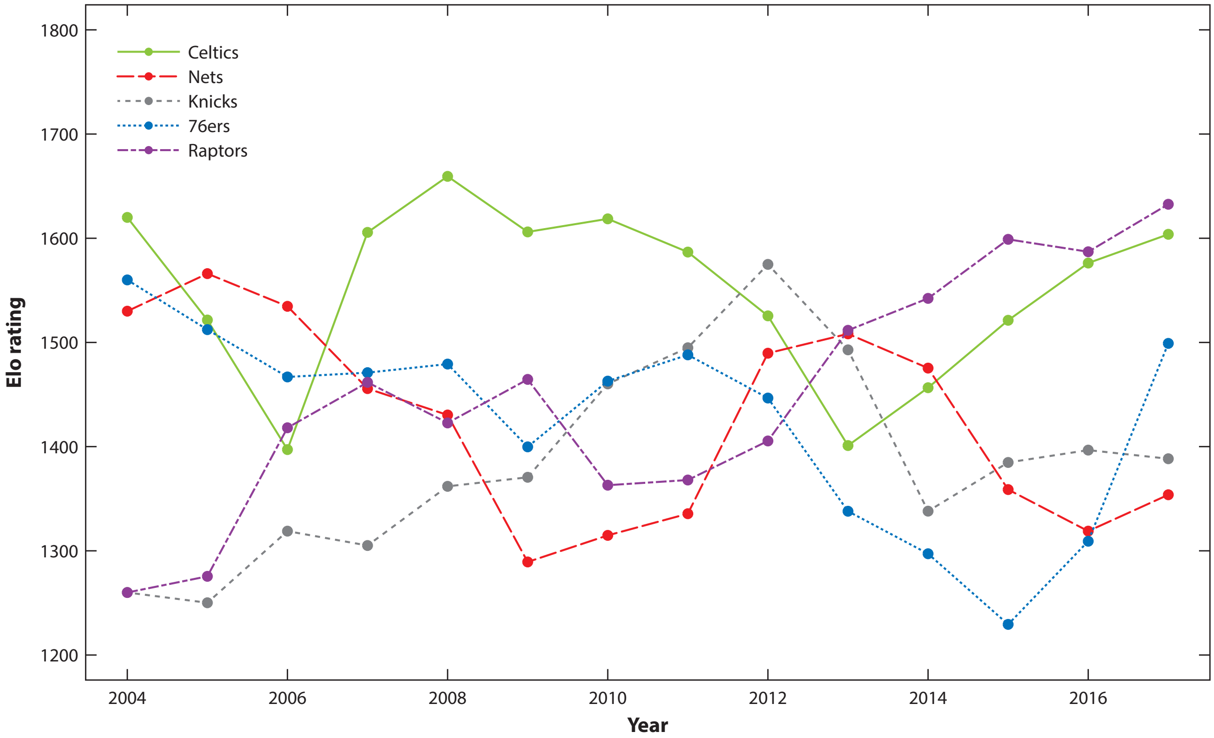

Figure 4

Elo ratings over time for the National Basketball Association Atlantic division, assuming optimal *K* factor.

Thurstone–Mosteller model due to Glenn & David (1960) mentioned in Section 2.1. Additionally, like Glickman (1999), TrueSkill assumes that player abilities evolve through time as a normal random walk. Like the Glicko and Glicko-2 systems, ratings are summarized as approximate normal posterior distributions, allowing for constructing interval estimates of strength. The main innovation of the TrueSkill system is its computational method, which relies on Bayesian updates through the expectation propagation algorithm (Minka 2001), an iterative algorithm that implements approximate message passing (Donoho et al. 2009). TrueSkill 2 differs from the original TrueSkill system by incorporating practical features, such as rating penalties for players who quit games before they have completed (an occurrence that is not uncommon in online games), and through simplifications to the computing algorithm (Minka et al. 2018).

Most implementations of TrueSkill and TrueSkill 2 are on Microsoft products, specifically on the Xbox network gaming platform. A Python implementation of the original TrueSkill can be obtained at **https://trueskill.org/**. As highlighted on Microsoft's TrueSkill website (Microsoft 2023), games such as Halo 3 and Forza Motorsport 7 have used the TrueSkill system. Microsoft has rolled out TrueSkill 2 for the games Gears of War 4 and Halo 5.

5.2. Universal Rating System

More elaborate rating systems have been developed to address idiosyncrasies not accounted for by these more basic rating systems. One such system for chess that was rolled out at the start of 2017 is called the Universal Rating System (URS) and has been the main rating system used in

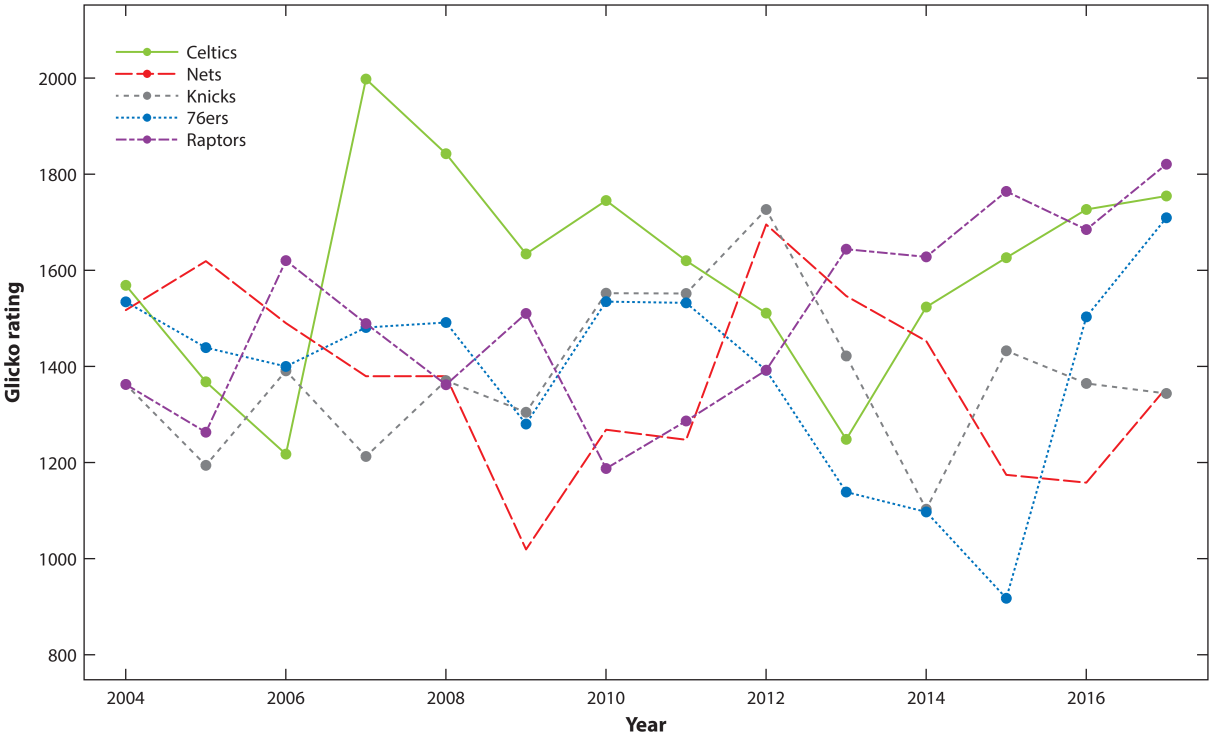

Figure 5

Glicko ratings over time for the National Basketball Association Atlantic division, assuming optimal innovation standard deviation σ.

tournament events run by the Grand Chess Tour (Grand Chess Tour 2021). We are not aware of any publicly available code that implements the URS.

One of the distinguishing features of the URS is that it addresses competitive chess games with different time limits for moves—that is, games that are played at different time controls. Unlike the Elo rating system, which assumes a single strength parameter per player, the URS assumes two strength parameters: one corresponding to games played at a speed of 60 moves within 5 minutes, and one to games played at a speed of 60 moves within 120 minutes. If a game is played at a time control other than 60 moves in 5 minutes and 60 moves in 120 minutes, the assumed strength parameter for the game is nonlinearly interpolated between these two extremes. The rating update then applies to these two strength parameters per player, with weights that depend on the time control of the game.

Additionally, the rating procedure for the URS is not a filtering update like the Elo or Glicko systems. Instead, the approach taken by the URS is to fit a paired comparison model (with the two strength parameters at different time controls mentioned above) using the most recent six years of game results, with likelihood terms that are time-weighted. Specifically, current games in the URS update are given full weight, and earlier games receive exponentially decaying weights in the likelihood function. Asymptotic standard error estimates of current ability can be obtained through standard likelihood-based methods, acknowledging the weights for game contributions in the likelihood. This approach to estimating current ability has been used by Dixon & Coles (1997) and McHale & Morton (2011) and discussed in a more general paired comparison setting in Cattelan et al. (2013).

5.3. Online Go

Various online Go servers have been in existence for years, offering players the opportunity to compete in games with opponents from around the world. One such server is the popular Kiseido Go Server (KGS), maintained by the American Go Foundation. The KGS system is based on the Bradley–Terry model and includes a novel method for updating ratings over time (Shubert 2006). The system essentially maximizes the Bradley–Terry log-likelihood, conditional on the opponents' current ratings, but includes game-specific weights that depend both on the time elapsed since the game was played and on the uncertainty of the opponents' ratings. Similar to the approaches by Dixon & Coles (1997) and McHale & Morton (2011), the weights decrease the longer ago the game was played. The uncertainty of an opponent's rating is determined from the second derivative of the log-likelihood from the most recent update of their rating, which can be viewed as an estimate of the reciprocal of the variance of the rating. One can argue that this approach underestimates the variance of the ratings, because the uncertainty in the opponents' ratings enters the calculation only through the weights and does not attempt to maximize the log-likelihood simultaneously across all the Bradley–Terry parameters. The KGS does not make available code for running its system or provide a site for running its code.

It is worth mentioning that, in the KGS system, a player's rating may change even when they have played no recent games. This can happen when reoptimizing the weighted log-likelihood where the player's opponents' ratings may have changed. This phenomenon rarely occurs in other modern rating systems.

6. DISCUSSION

Methods for assessing competitor strength or predicting game outcomes from head-to-head games have nearly a 100-year history, going back to the seminal papers by Zermelo (1928) and Thurstone (1927). The development of these models and their extensions to tie outcomes and to time-varying ability have enabled organizations to customize these approaches to rate competitors in large gaming organizations. The foundational methods for analyzing paired comparison outcomes have sparked a broad spectrum of innovation in further development. As technology and data science continue to evolve, there are plenty of avenues to explore, with the expectation of more nuanced approaches to predicting game outcomes.

The development of improved rating systems faces several challenges that have been known for years. One such problem involves the ability to rate competitors in isolated (or nearly isolated) pools. In gaming organizations, particularly with board games like chess, it is common for players to compete almost exclusively within limited geographic regions. Every once in a while, a small set of players choose to compete in national events, at which point they compete with opponents in different geographic regions. By not typically capturing the variability in strength estimation among players far apart on the player network, most rating models underestimate the level of uncertainty in game outcome probabilities between players coming from isolated pools. An area worthy of further development is rating methodology derived from models on networks. Success in applying network methods could help to quantify the uncertainty in outcomes between players.

Another common problem in rating systems based on paired comparisons is the potential for deflation or inflation over time. Deflation occurs when the distribution of computed ratings becomes lower over time despite player abilities (on average) not changing, while inflation analogously corresponds to rating increases without changes in ability. Elo (1978) explained that ratings have a natural tendency to deflate over time. He argued that, for chess (and by extension other games with large numbers of competitors), players typically begin competing at low strengths, get better over time with a corresponding increase in rating, and then leave the system. Using a rating system like Elo's that involves an equal exchange of rating points, the player essentially takes rating points out of the system once they retire from competition. Elo argued that this causes deflation because the rating points distributed among the pool of players is lower once the higher rated players retire. Glickman (1995) argued instead that the source of deflation is that the rating system has particular deficiencies that result in biased estimates for certain players. More concretely, a rating system may be able to update ratings of players whose abilities are slowly evolving over time. But if the rating system is unable to track quickly improving players, then their ratings will be lower than their abilities; as a consequence, they will defeat opponents more often than predicted by their ratings, thus causing a process in which the average of all players' ratings begins to decline.

These challenges highlight important open problems in the design and implementation of rating systems based on paired comparisons, and present rich areas for continued research. Future investigations could explore more sophisticated models that account for the nuances of player development, as well as system adjustments to more accurately reflect changes in player abilities. Additionally, developing methods to identify and correct for unintended behavior in rating systems could prevent the undue influence of rapidly improving players on the overall rating distribution, or in the difficulties associated with geographic clustering of competitors. By addressing these challenges, researchers can contribute to the creation of more stable, fair, and accurate rating systems that better serve competitive communities.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank two anonymous reviewers whose comments led to an improved manuscript.

LITERATURE CITED

ACF (Aust. Chess Fed.). 2013. *Ratings by-law*. By-Law Document, Aust. Chess Fed., Frankston South, Vic., Aust. **https://auschess.org.au/wp-content/uploads/2018/09/ACF-Ratings-By-Law.pdf**

Age of Empires DE Team. 2021. Updates to ranked team game ELO acquisition. *Age of Empires News Blog*, May 17. **https://www.ageofempires.com/news/rankedtg-update-5-2021**

Allebest E. 2018. Chess ratings – how they work. *Chess.com Help*, April 9. **https://www.chess.com/article/view/chess-ratings---how-they-work**

Araki K, Hirose Y, Komaki F. 2019. Paired comparison models with age effects modeled as piecewise quadratic splines. *Int. J. Forecast.* 35(2):733–40

Baker RD, McHale IG. 2014. A dynamic paired comparisons model: Who is the greatest tennis player? *Eur. J. Operat. Res.* 236(2):677–84

Baker RD, McHale IG. 2015. Deterministic evolution of strength in multiple comparisons models: Who is the greatest golfer? *Scand. J. Stat.* 42(1):180–96

Baker RD, McHale IG. 2017. An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player? *Eur. J. Operat. Res.* 258(1):328–33

Berrut JP, Floater MS, Klein G. 2011. Convergence rates of derivatives of a family of barycentric rational interpolants. *Appl. Numer. Math.* 61(9):989–1000

Bong H, Li W, Shrotriya S, Rinaldo A. 2020. Nonparametric estimation in the dynamic Bradley-Terry model. *Proc. Mach. Learn. Res.* 108:3317–26

Bradley RA, Terry ME. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39:324–45

Çakır G, Taifalos N, Davie C. 2024. Dota 2 ranks, MMR, and ranking system explained. *Dota 2 Blog*, July 25. **https://dotesports.com/dota-2/news/dota-2-mmr-and-ranking-system-explained**

Cattelan M. 2012. Models for paired comparison data: a review with emphasis on dependent data. *Stat. Sci.* 27(3):412–33

Cattelan M, Varin C, Firth D. 2013. Dynamic Bradley-Terry modelling of sports tournaments. *J. R. Stat. Soc. Ser. C* 62(1):135–50

David H. 1988. *The Method of Paired Comparisons*. London: Charles Griffin & Co.

Davidson RR. 1970. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Am. Stat. Assoc.* 65(329):317–28

Davidson RR, Beaver RJ. 1977. On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics* 33:693–702

Dewart N, Gillard J. 2019. Using Bradley-Terry models to analyze test match cricket. *IMA J. Manag. Math.* 30(2):187–207

Dixon MJ, Coles SG. 1997. Modelling association football scores and inefficiencies in the football betting market. *J. R. Stat. Soc. Ser. C* 46(2):265–80

Donoho DL, Maleki A, Montanari A. 2009. Message-passing algorithms for compressed sensing. *PNAS* 106(45):18914–19

Durbin J, Koopman SJ. 2012. *Time Series Analysis by State Space Methods*. Oxford, UK: Oxford Univ. Press

Elo AE. 1978. *The Rating of Chess Players, Past and Present*. New York: Arco

Fahrmeir L, Tutz G. 1994. Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Am. Stat. Assoc.* 89(428):1438–49

FICS (Free Internet Chess Serv.). 2008. Vek-splanation of the Glicko ratings system. *FICS Help File*. **https://www.freechess.org/Help/HelpFiles/glicko.html**

FIFA (Féd. Int. Footb. Assoc.). 2023a. *Men's ranking procedures*. Fact Sheet, FIFA, Zurich, Switz. **https://www.fifa.com/fifa-world-ranking/procedure-men**

FIFA (Féd. Int. Footb. Assoc.). 2023b. *Women's ranking procedures.* Fact Sheet, FIFA, Zurich, Switz. **https://www.fifa.com/fifa-world-ranking/procedure-women**

Floater MS, Hormann K. 2007. Barycentric rational interpolation with no poles and high rates of approximation. *Numer. Math.* 107:315–31

GGPoker. 2023. Spin & Gold ELO. **https://ggpoker.ca/poker-games/spin-gold-elo/**

Glenn WA, David HA. 1960. Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics* 16(1):86–109

Glickman ME. 1993. *Paired comparison models with time-varying parameters*. PhD Thesis, Dep. Stat., Harvard Univ., Cambridge, MA

Glickman ME. 1995. A comprehensive guide to chess ratings. *Am. Chess J.* 3(1):59–102

Glickman ME. 1999. Parameter estimation in large dynamic paired comparison experiments. *J. R. Stat. Soc. Ser. C* 48(3):377–94

Glickman ME. 2001. Dynamic paired comparison models with stochastic variances. *J. Appl. Stat.* 28(6):673–89

Glickman ME. 2022. Welcome to Glicko ratings. *Mark Glickman's World.* **http://www.glicko.net/glicko.html**

Glickman ME, Doan T. 2024. *The USCF rating system*. Tech. Doc., US Chess Fed., St. Louis, MO. **https://new.uschess.org/sites/default/files/media/documents/us_chess_rating_system_specs-2024-03-01.pdf**

Glickman ME, Stern HS. 2017. Estimating team strength in the NFL. In *Handbook of Statistical Methods and Analyses in Sports*, ed. J Albert, ME Glickman, TB Swartz, RH Koning, pp. 113–36. Boca Raton, FL: Chapman and Hall/CRC

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102(477):359–78

Good IJ. 1952. Rational decisions. *J. R. Stat. Soc. Ser. B* 14(1):107–14

Gorgi P, Koopman SJ, Lit R. 2019. The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *J. R. Stat. Soc. Ser. A* 182(4):1393–409

Grand Chess Tour. 2021. Universal Rating System$^{\mathrm{TM}}$. *Universal Rating System.* **http://universalrating.com/**

Herbrich R, Minka T, Graepel T. 2006. TrueSkill$^{\mathrm{TM}}$: a Bayesian skill rating system. In *NIPS'06: Proceedings of the 19th International Conference on Neural Information Processing Systems*, ed. B Schölkopf, JC Platt, T Hoffman, pp. 569–76. Cambridge, MA: MIT Press

Ingram M. 2021. How to extend Elo: A Bayesian perspective. *J. Quant. Anal. Sports* 17(3):203–19

Keith T. 2019. Backgammon FAQ: ratings. *Backgammon Galore*. **https://www.bkgm.com/faq/Ratings.html**

Knorr-Held L. 2000. Dynamic rating of sports teams. *J. R. Stat. Soc. Ser. D* 49(2):261–76

Koehler KJ, Ridpath H. 1982. An application of a biased version of the Bradley-Terry-Luce model to professional basketball results. *J. Math. Psychol.* 25(3):187–205

Krese B, Štrumbelj E. 2021. A Bayesian approach to time-varying latent strengths in pairwise comparisons. *PLOS ONE* 16(5):e0251945

League of Legends Wiki. 2022. Elo rating system. *League of Legends Wiki*. **https://leagueoflegends.fandom.com/wiki/Elo_rating_system**

Lichess.org. 2023. Frequently asked questions. *Lichess.org*. **https://lichess.org/faq#ratings**

Luce RD. 1959. *Individual Choice Behavior a Theoretical Analysis*. John Wiley and Sons

McHale I, Morton A. 2011. A Bradley-Terry type model for forecasting tennis match results. *Int. J. Forecast.* 27(2):619–30

Medado D. 2021. CS:GO ranks explained 2021 – how ranking system works, tips for good rank, complete guide. *AFK Gaming*. **https://afkgaming.com/csgo/guide/6825-csgo-ranks-explained-2021-how-ranking-system-works-tips-for-good-rank-complete-guide**

Microsoft. 2023. TrueSkill$^{\mathrm{TM}}$ ranking system. *Microsoft Research*. **https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/**

Minka T, Cleven R, Zaykov Y. 2018. *TrueSkill 2: an improved Bayesian skill rating system.* Tech. Rep. MSR-TR-2018-8, Microsoft, Redmond, WA

Minka TP. 2001. *A family of algorithms for approximate Bayesian inference*. PhD Thesis, MIT, Cambridge, MA

Morgulev E, Azar OH, Lidor R. 2018. Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* 5:213–22

Mosteller F. 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* 16(1):3–9

Noek A. 2017. OGS has a new Glicko-2 based rating system! *Online Go*. **https://forums.online-go.com/t/ogs-has-a-new-glicko-2-based-rating-system-**

Phelan GC, Whelan JT. 2018. Hierarchical Bayesian Bradley-Terry for applications in Major League Baseball. *Math. Appl* 7:71–84

Plackett RL. 1975. The analysis of permutations. *Appl. Stat.* 24(2):193–202

Rao PV, Kupper LL. 1967. Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *J. Am. Stat. Assoc.* 62(317):194–204

Sadasivan G. 1983. Within-pair order effects in paired comparisons. *Stud. Sci. Math. Hung.* 18:229–38

Scarf P, Rangel JS Jr. 2017. Models for outcomes of soccer matches. In *Handbook of Statistical Methods and Analyses in Sports*, ed. J Albert, ME Glickman, TB Swartz, RH Koning, pp. 357–70. Boca Raton, FL: Chapman and Hall/CRC

Shubert B. 2006. Rating system math. *KGS Go Server*. **https://www.gokgs.com/help/rmath.html**

Spanias D, Knottenbelt WJ. 2013. Predicting the outcomes of tennis matches using a low-level point model. *IMA J. Manag. Math.* 24(3):311–20

Stephenson A, Sonas J. 2020. PlayerRatings: dynamic updating methods for player ratings estimation. *R Package*, version 1.1-0. **https://CRAN.R-project.org/package=PlayerRatings**

Szymanski S. 2020. Sport analytics: science or alchemy? *Kinesiol. Rev.* 9:57–63

Taylor WJ. 1945. Method of Lagrangian curvilinear interpolation. *J. Res. Natl. Bureau Stand.* 35(2):151–55

Thurstone LL. 1927. A law of comparative judgment. *Psychol. Rev.* 34:273–86

Tsokos A, Narayanan S, Kosmidis I, Baio G, Cucuringu M, et al. 2019. Modeling outcomes of soccer matches. *Mach. Learn.* 108:77–95

USA Pickleball. 2018. How does the ELO system work? *USA Pickleball*. **https://usapickleball.org/ufaqs/how-does-the-elo-system-work**

Veček N, Črepinšek M, Mernik M, Hrnčič D. 2014. A comparison between different chess rating systems for ranking evolutionary algorithms. In *2014 Federated Conference on Computer Science and Information Systems*, pp. 511–18. Piscataway, NJ: IEEE

Watanabe NM, Shapiro S, Drayer J. 2021. Big data and analytics in sport management. *J. Sport Manag.* 35(3):197–202

West M, Harrison PJ, Migon HS. 1985. Dynamic generalized linear models and Bayesian forecasting. *J. Am. Stat. Assoc.* 80(389):73–83

Whelan JT, Klein JE. 2021. Bradley-Terry modeling with multiple game outcomes with applications to college hockey. arXiv:2112.01267 [stat.AP]

Yue JC, Chou EP, Hsieh MH, Hsiao LC. 2022. A study of forecasting tennis matches via the Glicko model. *PLOS ONE* 17(4):e0266838

Zermelo E. 1928. The calculations of the results of a tournament as a maximum problem in the calculus of probabilities. *Math. Z.* 29:436–60