Brazilian Journal of Biometrics, **44**, 1–16 (2026)
doi:10.28951/bjb.v44i3.894

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

A chess rating system that accounts for draws and the advantage of playing as white

Júlio Sílvio de Sousa Bueno Filho[⋆,1] and Danilo Machado Pires[2]

[1]Statistics Department, Federal University of Lavras, Lavras, MG, Brazil.
[2]Institute of Applied Social Sciences, Federal University of Alfenas, Varginha, MG, Brazil.
[⋆]Corresponding author. Email: jssbueno@ufla.br

(Received: May 15, 2025; Revised: March 27, 2026; Accepted: May 7, 2026; Published: August 10, 2026)
Section Editor: Douglas Mateus da Silva

Abstract

We introduce a chess rating system based on a preference model that includes an explicit parameter for draws, modified to account for the advantage of playing white. Heuristics derived from the conditional maxima of the hierarchical Bayesian model yield straightforward updating formulas. To illustrate the system's potential for describing past performance and predicting game results, we present the updating process using a small synthetic example and data from a recent tournament featuring elite chess players. The results are discussed and compared with other systems. The proposed system is competitive both in describing past performance and in predicting outcomes. It is flexible and has the potential to be used either as a chess rating system or to evaluate other sports where draws are possible.

Keywords: Performance evaluation; Posterior Maxima; Preference models; Sports.

1. Introduction

Chess rating systems are methods used to evaluate the relative skill levels of chess players. These systems aim to assign a numerical rating to each player, which represents their expected performance against other opponents. Ratings provide a standardized way to compare players and determine their level of expertise. One widely used rating system in chess is the Elo system, implemented by FIDE (Fédération Internationale des Échecs), the international regulatory body for the sport.

The Elo system, named after its creator Arpad Elo, calculates ratings based on the outcomes of chess games. Each player begins with an initial rating that, after a period, is adjusted based on game results. If a player performs better than expected against an opponent, their rating increases.

Conversely, if a player underperforms against an opponent, their rating decreases. The magnitude of rating changes depends on the rating difference between players.

FIDE uses a modified version of the Elo system for its official chess ratings (hereinafter called FIDE-Elo). It takes into account various factors such as the time control of the game, the ratings of opponents, and a smoothing parameter (*k*-factor) to regulate rating changes. Higher *k*-factor values are assigned to players with no previous registration or only a few games played, allowing them to make rapid progress. As players accumulate more games and their ratings stabilize, the *k*-factor decreases to ensure more gradual rating changes.

The FIDE rating system has become a global standard, providing a consistent measure of chess players' skill levels across different countries and competitions, and even other sports (e.g., it has been used in association football, or soccer). It is used to seed players in tournaments, calculate performance norms, and determine title awards such as FIDE Master, International Master, and Grandmaster. In short, it has been instrumental in fostering competition and allowing players to track their progress and compare themselves to others in the chess community.

Although widely accepted and used, the FIDE-Elo system has some limitations. One notable issue is the lack of granularity in rating adjustments, particularly when players of vastly different skill levels face each other. Additionally, the Elo system assumes that the variability (or uncertainty) of ratings, and thus the *k*-factor, is the same for large groups of players. This may not accurately reflect progress or decline in relative strength ((Glickman & Jones, 1999; Stephenson & Sonas, 2016)). Sometimes, rapid rating changes simply cannot be detected. The FIDE-Elo system does not consider the quality of moves or the performance within individual games, only the overall outcome. It also does not account for the advantage of playing with white or black pieces. Being a version of the Bradley and Terry preference model (with no draws), FIDE-Elo ignores tournament-level differences that could imply different drawing chances. A draw is simply treated as half a win (or half a loss). Elo (1978) considered that all these effects factor out over a long series of games and produce relatively unbiased estimates.

Many authors have proposed changes to FIDE-Elo to enhance its descriptive precision and predictive ability. Among these proposals are those that consider the advantage of playing as white, and the relative relevance of old and new results in composing the current rating update. Sismanis (2010) presents a report on the winner of the Kaggle/Deloitte/FIDE Chess Rating Challenge of 2010. The main contribution of the winning proposal is a time-dependency factor that shrinks the *k*-factor as the player becomes more active and increases this factor when the player is inactive. Other proposals estimate rating variances, *k*-factors, or rating volatilities for game-to-game updates. Among many proposals, we highlight the Glicko systems 1 and 2 (Glickman, 1995; Glickman, 2012), which include standard errors for rating estimates and an adaptive *k*-factor. Both systems have been successfully implemented in the USCF (United States Chess Federation) and on playing websites (such as https://www.chess.com and https://lichess.org). Some parameters in those systems are still set by the implemented software and are not updated based on results. Stephenson & Sonas (2016), reporting the win of the Kaggle/Deloitte/FIDE Chess Rating Challenge of 2012, proposed a refinement of the Glicko systems using ideas from Sismanis (2010).

In a slightly different fashion, Pires & Bueno Filho (2020) presented a Bayesian version of a modified preference model by Davidson (1970) with explicit parameters for draws and white advantage. In this proposal, a sampling approximation of the joint posterior distribution for a large set of parameters (ratings, white advantages, drawing tendency, and a time-lapse parameter for down-weighting the likelihood of old games) was obtained for the whole dataset. The method provided a good description of past performance and had good predictive properties. The drawback was the need to generate large samples from the joint posterior probabilities, which is impractical to do for each new tournament or game. In this paper, we present a chess rating system and update formulae that build on that approach. Model heuristics are derived from the approximation of the conditional maxima of the hierarchical Bayesian model.

In the following sections, we describe the rating system and the updating formulae. Some properties of the system are discussed using examples from games between players of typical rating levels, as well as from an important elite chess tournament held in January 2023.

2. Rating Systems

In this section, we discuss differences among existing rating systems and our new proposal. We base our discussion on the rating systems presented by Elo (1978) and Glickman (1995) and Glickman (2012). Given that there is sufficient evidence that learning curves and outcome probabilities depend on rating differences, which should be modeled as limited growth curves or (cumulative) distribution functions, we do not discuss linear systems such as Sonas (2005).

2.1 FIDE-Elo system

Elo (1978) assumed that rating differences have a Gaussian distribution with a mean of 0 and a standard deviation of $200\sqrt{2}$. He used a logarithmic approximation to yield the expected value of a win for the i^{th} player (as white) over the j^{th} player (as black) as follows:

$$E_{ij} = \frac{1}{1 + 10^{\frac{R_B - R_W}{400}}} \tag{1}$$

where R_B and R_W are the black and white rating parameters, respectively.

The updating formula for the Elo rating system is then:

$$R_{new} = R_{old} + k(W_O - W_E) \tag{2}$$

where R_{new} and R_{old} are the ratings after and before the games, respectively, and W_O and W_E are the observed and expected results from a set of games. For a single game involving the i^{th} player in (1), $W_E = E_{ij}$. For multiple games, we simply sum the expected results.

2.2 Glicko systems

The main contribution of Glickman (1995) was devising a rating system that follows a Bayesian framework and thus can include parameters for rating variability. Glicko 1 is a system that has been very successfully implemented (USCF, Chess.com, etc.). The model, however, is an adaptation of the same preference model (with no ties). Other parameters could be updated but are not usually implemented. Updating steps can be executed for each game or in a block of games from a given period, with all parameters updated simultaneously.

Glickman (2012) introduced an additional parameter for rating volatility, which enhances the variance of ratings due to inactivity. The algorithms are similar to those of the original system. Both systems have been successfully implemented by chess federations and websites for peer-to-peer gaming (Chess.com, 2023; Nolan & Scognamillo, 2021; Glickman & Doan, 2010).

2.3 Bayesian prediction system (BP): basic modeling

The preference model of Davidson (1970) is our basis for developing a new set of updating formulae for predicting results and rating chess players. Our extension devises the expected probability for each outcome for the game of the i^{th} player playing white against the j^{th} player playing black as follows:

$$\pi_{ij} = P(white\ wins) = \frac{e^{\gamma_i+\delta}}{e^{\gamma_i+\delta} + e^{\gamma_j-\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j}{2}}}$$

$$\pi_{ij0} = P(draw) = \frac{e^{\lambda+\frac{\gamma_i+\gamma_j}{2}}}{e^{\gamma_i+\delta} + e^{\gamma_j-\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j}{2}}} \quad (3)$$

$$\pi_{ji} = P(black\ wins) = \frac{e^{\gamma_j-\delta}}{e^{\gamma_i+\delta} + e^{\gamma_j-\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j}{2}}}$$

We used the transformation $\gamma_j = [\ln(10)/400]R_j = qR_j, \ \approx 0.005756R_j$ (and the respective back-transformation $R_j = \gamma_j/q \approx 173.7178\gamma_j$) to link parameters in the scale of our model to the commonly used FIDE-Elo system, in order to present ratings on a meaningful scale for chess enthusiasts.

The choice of numerators gives the proportions for the concentration parameters of a Dirichlet process for the multinomial distribution with outcomes $y = 0$, $y = 0.5$, and $y = 1$. Using δ for both π_{ij} and π_{ji} is arbitrary and allows for a cleaner presentation of π_{ij0} as a function not only of an extra parameter λ but also of the players' strengths. The rationale is that the stronger the players are, the larger the probability of a draw.

The λ parameter in these equations is a measure of the propensity to draw, and the δ parameter represents a rating boost from playing as white.

Joe (1990) provided a compelling rationale against using the formulation with explicit draws in the context of small datasets, citing significant variability of the resulting parameter across players. This phenomenon likely contributed to the near absence of such models in the chess rating literature. However, recent work by Pires & Bueno Filho (2020) provides evidence suggesting that, for large datasets, the inclusion of draws typically yields superior outcomes. Consequently, we found it worthwhile to reexamine the model.

In the subsequent section, we describe the modeling and the framework of the updating process. Additionally, we delineate scenarios in which the estimation of λ and σ may warrant different methods and tournament formats.

2.3.1 Prior distributions specification for BP

We chose to use Gaussian prior distributions for γ_i, γ_j, δ, and λ as follows: $\gamma_i|\sigma_A^2 \sim N(\gamma_{A0}, \sigma_A^2)$, $\gamma_j|\sigma_B^2 \sim N(\gamma_{B0}, \sigma_B^2)$, $\lambda|\sigma_\lambda^2 \sim N(\lambda_0, \sigma_\lambda^2)$, and $\delta|\sigma_\delta^2 \sim N(\delta_0, \sigma_\delta^2)$. The prior distributions adopted for the variance parameters are inverse-scaled chi-squared, with conjugacy to Gaussian algebraic forms, as follows: $\sigma_\delta^2 \sim \chi^{-2}_{scaled}(\nu_\delta, S_\delta^2)$, $\sigma_\lambda^2 \sim \chi^{-2}_{scaled}(\nu_\lambda, S_\lambda^2)$, $\sigma_A^2 \sim \chi^{-2}_{scaled}(\nu_A, S_A^2)$, $\sigma_B^2 \sim \chi^{-2}_{scaled}(\nu_B, S_B^2)$.

In this specification, we need to set values for the hyperparameters for degrees of freedom (ν_δ, ν_A, ν_B, and ν_λ) and for scaling factors (S_δ^2, S_A^2, S_B^2, and S_λ^2).

The prior distributions for the rating standard deviation are truncated for numerical stability at a maximum value of 350 FIDE-Elo (or 2.015 on the standardized scale). This truncation is needed for small examples with no previous learning of the parameter distribution and potential outliers (or highly unexpected results). This is done mainly to avoid possible discrepancies in interpreting the resulting scale, as individual players can have their σ_A^2 or σ_B^2 much larger than specified by FIDE-Elo.

As an example, the explicit algebraic form of the joint prior density distribution for γ_i is given, for a player $i = A$, as follows:

$$f(\gamma_A, \sigma^2_A) = (2\pi\sigma^2_{A0})^{-\frac{1}{2}} e^{-\frac{(\gamma_A-\gamma_{A0})^2}{2\sigma^2_{A0}}} \times \frac{(S^2_A)^{\nu_A}}{\Gamma(\nu_A)} (\sigma^2_A)^{-(\nu_A+1)} \exp\left(-S^2_A/\sigma^2_A\right) \tag{4}$$

The joint posterior distribution includes products of these terms and the multinomial likelihoods, presented in the next section. Sampling from the posterior distribution would be a computationally intensive and non-deterministic procedure. The main idea of this paper is to present direct updating formulae for each result, based on a one-step approximation from the maximum posterior equations, in which the current values of the parameters are taken as known and the new values as the solutions.

2.3.2 Observables

Realizations include the names of players playing white and black pieces (as designated by the tournaments) and the respective game results recorded as 0, 0.5, and 1, corresponding to a defeat, a draw, and a win for the player playing white, respectively.

$Y_g = I_{y_{ij}}$ is an indicator vector with ones (1) for the result of the game and zeros (0) otherwise. Game g has the following possible results: white victory ($y_{ij} = 1$), draw ($y_{ij} = 0.5$), or black victory ($y_{ij} = 0$). The respective $Y_g = I_{y_{ij}}$ vectors are $(1, 0, 0)$, $(0, 1, 0)$, or $(0, 0, 1)$. For example, if Veselin Topalov wins with the black pieces against Viswanathan Anand, the recorded vector is $Y_g = (0, 0, 1)$. If Magnus Carlsen and Veselin Topalov draw a game, the record is $Y_g = (0, 1, 0)$.

The likelihood function is of the form:

$$L = \pi_{ij}^{I_{y_{ij}=1}} \pi_{ij0}^{I_{y_{ij}=0.5}} \pi_{ji}^{I_{y_{ji}=0}}. \tag{5}$$

The log-likelihood (ℓL_r) for each outcome (white wins, draw, or black wins) is given by: $\ell L_1 = \gamma_i + \delta - \log(D)$ if white wins, $\ell L_2 = \lambda + \frac{\gamma_i+\gamma_j}{2} - \log(D)$ if the game is a draw, and $\ell L_3 = \gamma_j - \delta - \log(D)$ if black wins. In each case, the denominator of the probabilities is:

$$\log(D) = \log\left(e^{\gamma_i+\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j}{2}} + e^{\gamma_j-\delta}\right)$$

The log of the joint posterior distribution (ℓP) is given by adding ℓL_r and ℓp, the log of the prior distribution from equation 4. Thus, we can find the critical points for each j^{th} parameter as solutions of $\frac{\partial p(\theta|Y)}{\partial\theta_j} = 0 \Rightarrow \frac{\partial \log[p(\theta|Y)]}{\partial\theta_j} = \frac{\partial(\ell L_r+\ell p)}{\partial\theta_j} = 0$.

This well-known specification results in the log-concave posterior distribution

$$\ell P = log\left[p\left(\gamma_i, \gamma_j, \lambda, \delta, \sigma^2_{i0}, \sigma^2_{j0}, \sigma^2_{\lambda 0}, \sigma^2_{\delta 0} \mid Y_g\right)\right].$$

Thus, the critical point that solves for $\frac{\partial \log(D)}{\partial\gamma_i} = \pi_{ij} + \frac{\pi_{ij0}}{2}$ represents the global posterior maximum estimate and it is the main result of our rating updating idea.

In this process, the derivative of $\log(D)$ is needed for all three likelihoods and all parameters. We can describe these derivatives as direct functions of the probabilities in the model, as follows:

$$\frac{\partial \log(D)}{\partial\gamma_i} = \pi_{ij} + \frac{\pi_{ij0}}{2} \tag{6}$$

$$\frac{\partial \log(D)}{\partial\gamma_j} = \pi_{ji} + \frac{\pi_{ij0}}{2} \tag{7}$$

$$\frac{\partial \log(D)}{\partial \lambda} = \pi_{ij0} \tag{8}$$

$$\frac{\partial \log(D)}{\partial \delta} = \pi_{ij} - \pi_{ji} \tag{9}$$

2.3.3 Algorithm for updating formulae

The key step in the algorithm's heuristics is to consider known values of the parameters as point priors. In this way, the updates are steps toward values that maximize the joint posterior distribution. We implemented the following algorithm:

1) arrange games in chronological order;
2) set arbitrary initial values for all parameters;
3) run the updating formulae for γ_i and γ_j for each game result;
4) run updating formulae for δ and λ;
5) run updating formulae for variance parameters;
6) calculate predictions for subsequent games after pairings are declared.

Note that this recipe is flexible, and especially for steps 4 and 5, we could keep values fixed (making annual lists, for instance). On the other hand, in a tournament, updates could be performed at each round. Choosing a monthly period results in a more stable procedure and better suits advertising purposes, as FIDE or other federations would keep periodic records. This should be illustrated with larger datasets, but that is beyond the scope of this paper. Our main point is to have a set of clear rules for updating rating parameters from each possible outcome of a given game, as is currently done with the FIDE-Elo system, but with more descriptive features.

Note also that this resembles more general posterior approximation methods that can be used for a quick updating process when the full posterior distribution or summaries thereof are computationally intensive Marques (2018) and Lew *et al.* (2023).

2.3.4 Bayes prediction system updating formulae

In Table 1, we present updating formulae for each possible result of a single game between players with previously estimated (old) parameters. The change for game k can be denoted Δ_k and is a function of the prior directional variance and the difference between the observed and expected results.

$$\gamma_i^{new} = \gamma_i^{old} + \Delta_k = \gamma_i^{old} + \sigma_{i0}^2 \left(Y_g - E_0[Y_g]\right)$$

This can be summed over n games in a given period, so that the update $\sum_{k=1}^{n} \Delta_k$ is produced, taking current ratings as fixed. Monthly updates, for instance, are useful for later comparisons with other updating systems on actual data.

As an example, to update γ_i for a single game (k) based on a defeat with white pieces, we use Δ_k from the first row and the "Black wins" column of the table:

$$\gamma_i^{new} = \gamma_i^{old} + \Delta_k = \gamma_i^{old} + \sigma_{i0}^2 \left[0 - \left(\pi_{ij(k)} + \frac{1}{2}\pi_{ij0(k)}\right)\right]$$

Table 1. Tabulated values represent changes Δ_k from current (old) values of the parameters $\gamma_{i0}, \gamma_{j0}, \delta_0$, and λ_0. Updating formulae for linear parameters from the Bayes prediction rating system: ratings of white and black players (γ_i and γ_j), white advantage (δ), and propensity to draw (λ)

	Black wins $y = 0$	Draw $y = 0.5$	White wins $y = 1$
γ_i	$-\sigma^2_{i0}\left(\pi_{ij(k)} + \frac{1}{2}\pi_{ij0(k)}\right)$	$\sigma^2_{i0}\left(\frac{1}{2} - \pi_{ij(k)} - \frac{1}{2}\pi_{ij0(k)}\right)$	$\sigma^2_{i0}\left(1 - \pi_{ij(k)} - \frac{1}{2}\pi_{ij0(k)}\right)$
γ_j	$\sigma^2_{j0}\left(1 - \pi_{ji(k)} - \frac{1}{2}\pi_{ij0(k)}\right)$	$\sigma^2_{j0}\left(\frac{1}{2} - \pi_{ji(k)} - \frac{1}{2}\pi_{ij0(k)}\right)$	$-\sigma^2_{j0}\left(\pi_{ji(k)} + \frac{1}{2}\pi_{ij0(k)}\right)$
δ	$-\sigma^2_{\delta 0}\left(1 + \pi_{ij(k)} - \pi_{ji(k)}\right)$	$-\sigma^2_{\delta 0}\left(\pi_{ij(k)} - \pi_{ji(k)}\right)$	$\sigma^2_{\delta 0}\left(1 - \pi_{ij(k)} + \pi_{ji(k)}\right)$
λ	$-\sigma^2_{\lambda 0}\pi_{ij0(k)}$	$\sigma^2_{\lambda 0}(1 - \pi_{ij0(k)})$	$-\sigma^2_{\lambda 0}\pi_{ij0(k)}$

Updating formulae for variance parameters were derived from the same one-step maximization of the joint full posterior. Within brackets are the squares of the updating changes Δ_k from individual games, as presented in Table 1. Squaring each of those Δ_k updates and using them in equation (10) updates the variance parameters for ratings, white advantage, and propensity to draw (σ^2_l, where l represents $\gamma_i, \gamma_j, \lambda$, and δ). If one wants to implement updates for each game, simply set $n = 1$.

$$\hat{\sigma}^2_l = \frac{\nu_l S^2_l + \sum_{k=1}^{n} \Delta^2_k}{\nu_l + n + 2} \tag{10}$$

2.4 Posterior Improvement Guarantee

For the full posterior maximization no extra guarantee is needed, as the updating directly searches for the maximum of the posterior of a single game. However, for tournament and marketing purposes, it is more convenient to perform the updating process using longer periods and larger game sets for global parameters, and not after every game.

We must verify that this block-wise procedure does not decrease the overall joint posterior density.

Since the conditional log-posterior is strictly concave in (γ_i, γ_j) the maximum is unique and global within the admissible domain. Thus, by definition of the maximum, we have:

$$p\left(\gamma_i^*, \gamma_j^* \mid \lambda, \delta, \sigma^2_{i0}, \sigma^2_{j0}, \sigma^2_{\lambda 0}, \sigma^2_{\delta 0}, Y_g\right) \geq p\left(\gamma_{i0}, \gamma_{j0} \mid \lambda, \delta, \sigma^2_{i0}, \sigma^2_{j0}, \sigma^2_{\lambda 0}, \sigma^2_{\delta 0}, Y_g\right).$$

This block-wise maximization is a particular instance of the ECM algorithm (Meng & Rubin, 1993), which extends the classical EM algorithm (Dempster *et al.*, 1977). The EM theory guarantees that each conditional maximization step produces a non-decrease in the objective function. Extending this result to maximum a posteriori (MAP) estimation is straightforward as the addition of any constants (from the log-prior) does not affect the monotonicity inequality.

Therefore, the proposed block-wise updating scheme generates a sequence of parameter estimates for which the joint posterior density is non-decreasing at each step. Convergence to a local maximum follows under standard regularity conditions.

2.5 Methods and exemplification

In the remaining sections, we will compare several applications of the BP system with FIDE-Elo and, eventually, Glicko systems. To evaluate the surprising effect of a single result based on previously known parameters, we used an approximate χ^2_ν statistic given by $-2\log P(Y_g \mid \theta_g)$, where

Y_g is the observed result and θ_g is a vector containing all parameters in the model that can be updated by the specific game. For FIDE-Elo and Glicko systems, it is more natural to consider only two outcomes (and thus $\nu = 1$). For our model, which contemplates draws, the appropriate choice for the number of degrees of freedom would be $\nu = 2$. To approximate the likelihood of a given game after it occurred, we used $-2 \log P(Y_g \mid \theta_g^*)$, updating the values of the parameters in θ_g^*. It is expected that post-game values will be smaller than their pre-game counterparts as a result of the updating process. Note that pre-game statistics can be converted to a log-odds score and may eventually offer some guidance for betting decisions, while post-game statistics are more suitable for evaluating specific characteristics of a game or tournament that has just occurred (post-game or post-tournament).

To ensure a fair comparison, we performed two alternative adaptations: adapting our BP model to only two outcomes, and adapting FIDE-Elo and Glicko to three outcomes. Of course, it is expected that each method will show some superiority in its originally designed scenario. To consider only two results in BP, we define:

$$\pi = P(\textit{white win}) = (\pi_{ij} + \pi_{ij0})/2$$

and

$$P(\textit{white loss}) = 1 - \pi.$$

Thus

$$\log(P(y)) = y \log(\pi) + (1 - y) \log(1 - \pi).$$

For the models considering only two results, the formula for the probabilities of draws was obtained by letting $\pi = P(\textit{white win})$ as follows:

$$C = \pi + \sqrt{\pi(1 - \pi)} + (1 - \pi).$$

Thus

$$\pi_{ij} = \frac{\pi}{C},$$

$$\pi_{ij0} = \frac{\sqrt{\pi(1 - \pi)}}{C},$$

$$\pi_{ji} = \frac{1 - \pi}{C}.$$

And for $I_Y = (y_w, y_d, y_b)$, representing indicators for white win, draw, and black win, we obtain:

$$\log(P(y)) = y_w \log(\pi_{ij}) + y_d \log(\pi_{ij0}) + y_b \log(\pi_{ji}).$$

2.5.1 A synthetic example

To illustrate the updating process, suppose players A, B, C, D, E, and F play six games (say, $A \times B$, $C \times D$, and $E \times F$, and three more with reversed colors) as presented in Table 2. This example was designed to illustrate the differences in updates resulting from different colors and player strengths. Current ratings are known (on the FIDE-Elo scale) before each game. The probabilities for each result are computed using the BP model, with arbitrary values $\lambda_0 = 20q$ and $\delta_0 = 30q$. For these games, we present direct contributions to the rating updates for all parameters. For practical implementations, aggregate values could be used, and more than one game could be considered in the updating process.

Table 2. Probabilities (in %) for each chess game result using the proposed system (BP). Games 4 to 6 involve the same player strengths as games 1 to 3 but with reversed colors. Games 7 to 9 are between players of equal strength and reflect the advantage of playing white

	Player-Rating		Probability of the outcome (%)		
Game	White	Black	Black wins	Draw	White wins
1	A-1400	B-1800	63.98	26.98	09.04
2	C-2100	D-2400	55.14	31.01	13.85
3	E-2600	F-2800	45.53	34.14	20.34
4	1800-B	1400-A	05.17	21.80	73.03
5	2400-D	2100-C	08.26	26.12	65.62
6	2800-F	2600-E	12.76	30.25	56.99
7	1600-G	1600-H	26.69	35.60	37.71
8	2250-I	2250-J	26.69	35.60	37.71
9	2700-L	2700-M	26.69	35.60	37.71

In Table 3 we present the updated values for all parameters in each game, for each possible result. Note that in no case were the estimated variances close to 350^2. This value is an upper limit suggested by Glickman and Jones (1999) and adopted by Stephenson and Sonas (2020) to avoid unstable updates in their methods. In our case, no truncation was needed, even when starting from large initial variances.

The most important feature of this table is the asymmetric updating for pairs of players with the same strength when colors are reversed. This asymmetry is not present in any current updating system and has the potential to improve both performance description and predictions.

Table 3. Updated estimates and standard deviations (in the FIDE-Elo scale) for all parameters, considering all possible results from the games in Table 2, with $\delta = 30$, $\lambda = 20$, $\sigma_A = \sigma_B = 60$, $\sigma_\delta = 25$, and $\sigma_\lambda = 30$

Game	Parameter	Black wins		Draw		White wins	
		Estimate	SD	Estimate	SD	Estimate	SD
1	γ_i ($\gamma_i^{old} = 1400$)	1395.3	27.7	1405.7	27.7	1416.1	28.1
	γ_j ($\gamma_j^{old} = 1800$)	1804.7	27.7	1794.3	27.7	1783.9	28.1
	δ	26.5	41.6	30.6	41.6	34.2	41.6
	λ	18.6	27.3	23.8	27.7	18.6	27.7
2	γ_i ($\gamma_i^{old} = 2100$)	2093.9	27.7	2104.3	27.7	2114.6	28.0
	γ_j ($\gamma_j^{old} = 2400$)	2406.0	27.7	2395.7	27.7	2385.4	28.0
	δ	26.4	41.6	30.6	41.6	34.2	41.6
	λ	18.4	27.7	23.6	27.7	18.4	27.7
3	γ_i ($\gamma_i^{old} = 2600$)	2592.2	27.8	2602.6	27.7	2612.9	27.9
	γ_j ($\gamma_j^{old} = 2800$)	2807.7	27.8	2797.4	27.7	2787.9	27.9
	δ	27.3	41.6	30.9	41.6	34.5	41.6
	λ	18.2	27.7	23.4	27.7	18.2	27.7
4	γ_i ($\gamma_i^{old} = 1800$)	1782.6	28.2	1792.9	27.8	1803.3	27.7
	γ_j ($\gamma_j^{old} = 1400$)	1417.4	28.2	1407.0	27.8	1396.7	27.7
	δ	23.9	41.6	27.5	41.6	31.2	41.6
	λ	18.8	27.7	24.0	27.7	18.9	27.7
5	γ_i ($\gamma_i^{old} = 2400$)	2383.7	28.1	2394.0	27.8	2404.4	27.7
	γ_j ($\gamma_j^{old} = 2100$)	2116.3	28.1	2105.9	27.8	2095.6	27.7
	δ	24.3	41.6	27.9	41.6	31.5	41.6
	λ	18.6	27.7	23.8	27.7	18.6	27.7
6	γ_i ($\gamma_i^{old} = 2800$)	2785.0	28.0	2795.4	27.7	2805.8	27.7
	γ_j ($\gamma_j^{old} = 2600$)	2614.9	28.0	2604.6	27.7	2594.2	27.7
	δ	24.8	41.6	28.4	41.6	32.0	41.6
	λ	18.4	27.7	23.6	27.7	18.4	27.7
7	γ_i ($\gamma_i^{old} = 1600$)	1588.5	27.9	1598.8	27.7	1609.2	27.8
	γ_j ($\gamma_j^{old} = 1600$)	1611.5	27.9	1601.1	27.7	1590.8	27.8
	δ	26.0	41.6	29.6	41.6	33.2	41.6
	λ	18.2	27.7	23.3	27.7	18.1	27.7
8	γ_i ($\gamma_i^{old} = 2250$)	2238.5	27.9	2248.9	27.7	2259.2	27.8
	γ_j ($\gamma_j^{old} = 2250$)	2261.5	27.9	2251.1	27.7	2240.8	27.8
	δ	26.0	41.6	29.6	41.6	33.2	41.6
	λ	18.2	27.7	23.3	27.7	18.2	27.7
9	γ_i ($\gamma_i^{old} = 2700$)	2688.5	27.9	2698.9	27.7	2709.2	27.8
	γ_j ($\gamma_j^{old} = 2700$)	2711.5	27.9	2701.1	27.7	2690.8	27.8
	δ	26.0	41.6	29.6	41.6	33.2	41.6
	λ	18.1	27.7	23.3	27.7	18.1	27.7

3. Results and Discussion

3.1 A recent elite tournament application

Our BP updating formulae system was evaluated and compared to other well-known systems using an elite chess tournament: the Tata Steel Chess Masters, held in January 2023 (*Tata Steel Chess Tournament* (2023)). Table 4 presents the players, initial ratings, results for each game, and final scores. All tournament details can be found at https://chess-results.com/tnr719076.aspx?lan=10&art=2&flag=30.

Table 4. Cross table after 13 rounds. Players are listed in final classification order

	Player	Rating	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Score
1	Giri, A.	2764		.5	1	.5	.5	.5	.5	1	.5	.5	1	1	.5	.5	8.5
2	Abdusattorov, N.	2713	.5		1	.5	.5	1	.5	1	.5	0	.5	.5	.5	1	8
3	Carlsen, M.	2859	0	0		.5	1	1	.5	1	.5	.5	.5	.5	1	1	8
4	So, W.	2760	.5	.5	.5		.5	.5	.5	.5	.5	.5	.5	1	1	.5	7.5
5	Caruana, F.	2766	.5	.5	0	.5		.5	.5	.5	.5	1	.5	1	.5	.5	7
6	Maghsoodloo, P.	2719	.5	0	0	.5	.5		1	.5	1	1	.5	0	.5	1	7
7	Aronian, L.	2735	.5	.5	.5	.5	.5	0		.5	.5	.5	.5	.5	1	.5	6.5
8	Rapport, R.	2740	0	0	0	.5	.5	.5	.5		1	.5	1	.5	.5	1	6.5
9	Rameshbabu, P.	2684	.5	.5	.5	.5	.5	0	.5	0		1	1	0	.5	.5	6
10	Van Foreest, J.	2681	.5	1	.5	.5	0	0	.5	.5	0		.5	.5	.5	1	6
11	Ding, L.	2811	0	.5	.5	.5	.5	.5	.5	0	0	.5		1	.5	.5	5.5
12	Gukesh, D.	2725	0	.5	.5	0	0	1	.5	.5	1	.5	0		.5	.5	5.5
13	Keymer, V.	2696	.5	.5	0	0	.5	.5	0	.5	.5	.5	.5	.5		.5	5
14	Erigaisi, A.	2722	.5	0	0	.5	.5	0	.5	0	.5	0	.5	.5	.5		4

This was a very interesting tournament, marking the (to date) last appearance of Norwegian GM Magnus Carlsen as FIDE world champion. It is noteworthy that he was beaten by the young Uzbek GM Nodirbek Abdusattorov, who had excellent results previously (second place, a 146-rating-point underdog), and by Dutch GM Anish Giri, who had the tournament of his life (first place, a 95-rating-point underdog).

A summary of the probabilities before each game for the most unexpected results is given in Table 5. Despite the media stir, none of these results were particularly surprising. Not even the first game reached the usual threshold of $p < 5\%$ to reject the null hypothesis that the parameters are adequate for this game. In these examples, there is no need to compute post-game surprise effects. Note also that the soon-to-be world champion, Chinese GM Ding Liren, won only one game, against the young Indian star GM Dommaraju Gukesh (an 86-rating-point underdog). Despite drawing with Carlsen in their game, Ding failed to exert his favoritism in many games.

Table 5. Probabilities ($\times 100\%$) for each result in selected games. Boldface indicates the actual results. The surprise effect of the result given the current rating is computed as $P(y \mid \theta) \approx p(\chi^2_{\nu=2} > \chi^2_c) \times 100\%$

	Player-Rating		Result			Surprise
Game	White	Black	0	0.5	1	
1	Carlsen, M.	Abdusattorov, N.	**15.86**	32.21	51.93	5.82
2	Giri, A.	Carlsen, M.	35.31	35.82	**28.87**	16.61
3	Carlsen, M.	Ding, L.	22.76	**34.85**	42.39	19.05
4	Gukesh, D.	Ding, L.	34.46	35.87	**29.67**	9.72

In Table 6 we present a direct comparison of FIDE-Elo with our BP system at the end of the tournament, using game-by-game updates. The actual FIDE-Elo system, as currently implemented, would use the entire tournament for updates. The updates for each player keep the previous month's values as a reference. For better comparison, we used whole-round updates for both systems. In this updating scheme, for a given player, game-to-game updates tend to have smaller impact than monthly updates. On average, one expects the same results over long periods. Monthly updates are quite practical for tournament organizers and advertising. Game-by-game updates tend to be more practical for online games. In this context, any fixed period lacks practical meaning.

Table 6. Comparison of rating updates and tournament performance description. Elo ratings used $k = 10$. BP used $\delta = 25$, $\sigma_\delta = 15$, $\lambda = 50$, $\sigma_\lambda = 15$. Initial values for BP ratings (γ_i) were the FIDE-Elo ratings, with $\sigma_{\gamma_i} = 60$

Tourn.	Player	FIDE-Elo			BP		
order		Δ_k	world	SD	Δ_k	world	SD
1	Giri, A.	14.51	3	4.41	8.51	3	27.74
2	Abdusattorov, N.	17.77	8	8.11	16.79	8	27.74
3	Carlsen, M.	-6.62	1	5.00	-1.51	1	27.74
4	So, W.	5.76	4	4.03	2.82	5	27.74
5	Caruana, F.	-0.92	5	2.49	0.14	4	27.74
6	Maghsoodloo, P.	9.28	9	3.60	2.11	9	27.74
7	Aronian, L.	0.64	7	2.77	2.47	6	27.74
8	Rapport, R.	0.48	6	3.86	-8.73	7	27.74
9	Rameshbabu, P.	5.10	13	4.65	-0.05	13	27.74
10	Van Foreest, J.	7.04	14	3.67	1.21	14	27.74
11	Ding, L.	-22.14	2	8.74	-0.30	2	27.74
12	Gukesh, D.	-5.40	10	4.23	-10.89	10	27.74
13	Keymer, V.	-5.00	12	3.31	-3.25	12	27.74
14	Erigaisi, A.	-20.46	11	7.75	-9.90	11	27.74

Prior values for the parameters were chosen to make our model (initial values and updates) similar to FIDE-Elo. Prior variances were kept large enough to allow for initial learning, as the dataset is small. Posterior values were $\sigma_{\gamma_i} \approx 27.7$, $\delta = 23.8$ with $\sigma_\delta = 34.4$, and $\lambda = 105.5$ with $\sigma_\lambda = 23.0$, indicating decreasing values for the posterior distributions. This is a potential indication of adequate model learning from the data. As usual in elite tournaments, there were many draws, which is reflected in the increasing λ parameter. Greater variations in rating from our model can be observed compared to the FIDE-Elo model.

In Figure 1, the upper graphs depict rating evolutions for the players with the four highest final ratings according to each system. In the bottom graphs, the same is done for the lowest-rated players. There was agreement in the ordering of the top three players by both systems, with a disagreement for fourth place. Note that the FIDE-Elo system was very favorable to Wesley So's performance, while ours was beneficial to Fabiano Caruana. This is largely because Caruana had a higher initial rating by six points, and his results did not contain major surprises. Wesley So was undefeated, and this performance is well rewarded in the FIDE-Elo system but not in our BP system, as many opportunities to gain a full point as white were missed. Rating loss for the world number one was also less pronounced in our system than in FIDE-Elo. In general, BP exhibits smoother variation (and with large datasets it is expected to have less intense variation) than FIDE-Elo, as also shown in Table 6.

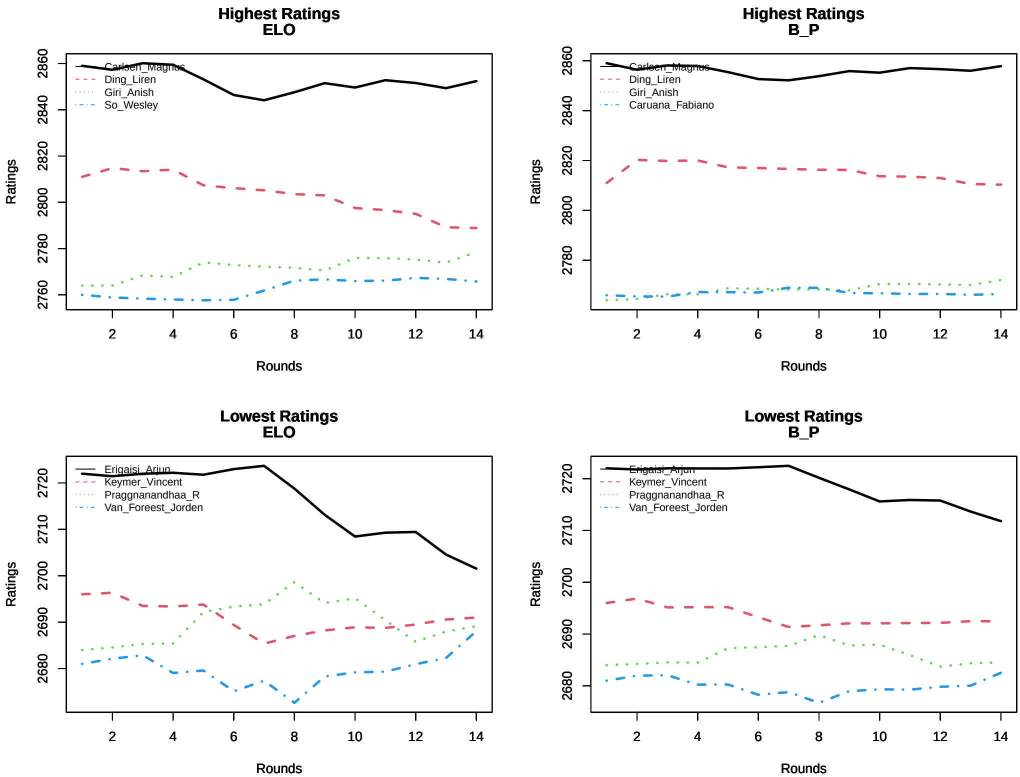

Figure 1. Evolution of the four highest-rated and four lowest-rated players using FIDE-Elo and the BP system.

In Figure 2, we present the average log-likelihood statistics for all results in each round. It is observed that the Bayesian Prediction (BP) rating system does not suffer from a slight upward bias when adapted to two outcomes, while still maintaining comparable descriptive and predictive characteristics. Conversely, when considering all possible outcomes, the BP model demonstrates significantly enhanced accuracy, making it a preferable choice. This observation strongly suggests that BP represents a viable improvement over FIDE-Elo and related models that do not account for draws.

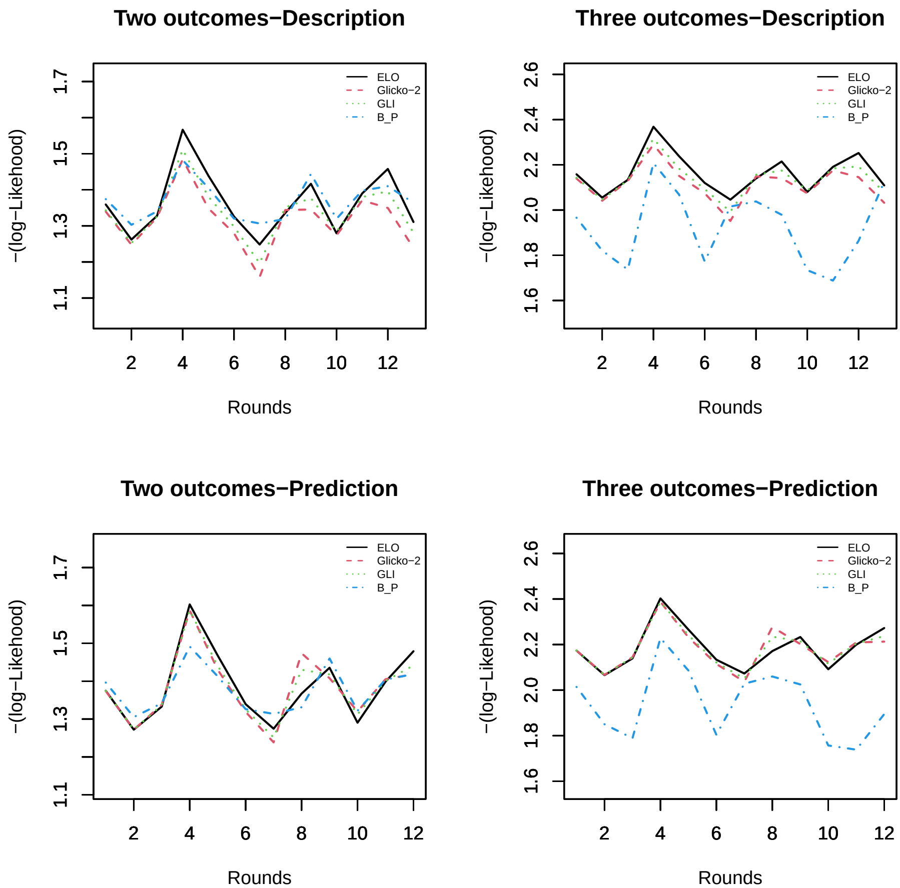

Figure 2. Average log-likelihood for four rating systems considering two and three outcomes. The upper graphs use post-game probabilities (for descriptive purposes). The lower graphs use pre-game probabilities (for predictive purposes).

3.2 Practical implications

In this specific case, we implemented parameter updates on a per-round basis. Nonetheless, this methodology may lack practicality when applied to real-world scenarios. Our system affords adaptability to support periodic updates, whether on an annual, tournament-centric, or round-specific schedule. Indeed, a static version of the foundational BP system model was implemented, utilizing various choices of parameter sets, and demonstrated promise when applied to extensive datasets (Pires & Bueno Filho (2020)). In that paper, the advantage of playing white was also modeled as individual-based, generating four parameters (γ_i and δ_i and their respective variances) for each player. It should be emphasized that the intricacies of these alternative update methodologies are beyond the scope of this paper, but some of the practical implications can be anticipated.

For each player, the parameters γ_i, σ_{γ_i} (and potentially δ_i and σ_{δ_i}) could be subject to monthly updates and public disclosure by the relevant federation. It may be pragmatic to consider collective parameters governing white advantage (δ and σ_δ) with less frequent updates, such as annually. Parameters related to the propensity for draws (λ and σ_λ) could be updated annually for tournaments within specific rating categories. For example, for a beginner's tournament with an average rating of $\bar{\gamma} = 1300$, we might use a lower λ value than for a master's tournament with an average rating of

$\bar{\gamma} = 2400$. Additionally, there may be cases where specific σ_λ values should be retained.

In summary, the adoption of such a new system could promote significant changes in how player ratings are managed and, consequently, how pairings are determined, accounting for venue-specific and tournament-round-specific situations. This could, for instance, enable the assignment of distinct probability sets for outcomes in pairings during the initial versus final rounds of specific master-level tournaments.

When Elo initially proposed his system, manual calculations using logarithmic tables were the norm; today, mobile applications fed by data stored on web platforms can instantly inform all players about specific pairing details. We anticipate that this advancement could open new avenues for promoting the sport and facilitating technical discussions with more informed insights.

4. Conclusions

In this paper, we have presented a new rating system and derived updating formulae from a Bayesian-inspired heuristic to approximate the posterior maxima. The system allows for directly predicting the probabilities of each outcome of a single game between two chess players (including draws), as well as using a batch of games for both prediction and rating updates. It is a flexible system that has the potential to be used either for chess ratings or to describe and predict performances in other sports with three possible outcomes (win, draw, or loss).

Acknowledgments

The authors would like to thank the funding agencies CNPq, CAPES, and FAPEMIG for supporting some of our related research, as well as the anonymous referees who pointed out clear improvements to this paper.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: BUENO FILHO, J. S. S.; PIRES, D. M. **Data curation:** BUENO FILHO, J. S. S.; PIRES, D. M. **Formal analysis:** BUENO FILHO, J. S. S.; PIRES, D. M. **Investigation:** BUENO FILHO, J. S. S.; PIRES, D. M. **Methodology:** BUENO FILHO, J. S. S.; PIRES, D. M. **Software:** BUENO FILHO, J. S. S.; PIRES, D. M. **Resources:** BUENO FILHO, J. S. S.; PIRES, D. M. **Supervision:** BUENO FILHO, J. S. S. **Validation:** BUENO FILHO, J. S. S.; PIRES, D. M. **Visualization:** BUENO FILHO, J. S. S.; PIRES, D. M. **Writing original draft:** PIRES, D. M. **Writing-review and editing:** BUENO FILHO, J. S. S.; PIRES, D. M.

Supplementary Material

R code for analysis: The R code used for implementing rating evaluations can be found at https://github.com/Arquivos-code/Baysian-predicition.git

Tata Steel 2023 data set: Data set used in the illustration of methods in Section 3. (TataSteel2023.txt file)

References

1. Chess.com. *Chess Terms* https://www.chess.com/terms/elo-rating-chess.
2. Davidson, R. R. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65,** 317–328 (1970).

3. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39,** 1–22 (1977).
4. Elo, A. E. *The rating of chessplayers, past and present* (Arco Pub., 1978).
5. Glickman, M. E. Example of the Glicko-2 system. *Boston University,* 1–6 (2012).
6. Glickman, M. E. The glicko system. *Boston University* **16,** 16–17 (1995).
7. Glickman, M. E. & Doan, T. The USCF rating system. *URL http://www.glicko.net/ratings/rating.system.pdf* (2010).
8. Glickman, M. E. & Jones, A. C. Rating the chess rating system. *CHANCE* **12,** 21–28 (1999).
9. Joe, H. Extended use of paired comparison models, with application to chess rankings. *Journal of the Royal Statistical Society Series C: Applied Statistics* **39,** 85–93 (1990).
10. Lew, A. K., Matheos, G., Tan Z.-X., T., Ghavamizadeh, M., Gothoskar, N., Russell, S. & Mansinghka, V. K. *Smcp3: Sequential monte carlo with probabilistic program proposals* in *International Conference on Artificial Intelligence and Statistics* (2023), 7061–7088.
11. Marques, R. A. G. *On Monte Carlo contributions for real-time probabilistic inference* eng. Oslo, 2018.
12. Meng, X.-L. & Rubin, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278 (1993).
13. Nolan, E. & Scognamillo, V. Online Chess Social Networks (2021).
14. Pires, D. M. & Bueno Filho, J. S. S. CAN ELO RATINGS BE IMPROVED? A CASE STUDY WITH ELITE CHESS PLAYERS. *Brazilian Journal of Biometrics* **38,** 483–505 (2020).
15. Sismanis, Y. How i won the" chess ratings-elo vs the rest of the world" competition. *arXiv preprint arXiv:1012.4571* (2010).
16. Sonas, J. *Chessmetrics* http://www.chessmetrics.com/cm/. 2005.
17. Stephenson, A. & Sonas, J. *Playerratings: Dynamic updating methods for player ratings estimation [Computer software manual]* 2016.
18. *Tata Steel Chess Tournament* https://tatasteelchess.com. May 2023.