CAN ELO RATINGS BE IMPROVED? A CASE STUDY WITH ELITE CHESS PLAYERS

Danilo Machado PIRES[1]
Júlio Sílvio de Sousa BUENO FILHO[2]

▪ ABSTRACT: Originally designed as a way to reflect past performance, chess ratings are now widely used to reflect players strength with many important aspects in tournament scheduling, advertising and premium shares. The ELO system has been officially adopted by World Chess Federation (FIDE). We used Bayesian analysis of actual data from elite chess players to fit parametric statistical models that could subsidize proposals for rating system improvement. Although most of the considered options are not new, since based on well known preference models, the use of a weighed likelihood function to emulate dynamic rating systems via Bayesian inference is novel. We compared descriptive ability using marginal likelihood based information criteria. Akaike information criterion was used to compare predictions. Many of the considered options improve on Elo ratings and there is strong evidence that dynamic models considering both white advantage and propensity to draws would result in more accurate systems.

▪ KEYWORDS: Bayesian inference; performance evaluation; preference models; sports.

1 Introduction

Chess is one of the most popular games in the world, being practiced by millions of people (formally or informally). There is a lot of literature on chess, thousands of books and magazines, websites and data banks to retrieve games played from 15 century on and many sources of information on tournaments and players history. Early introduction of rating systems to estimate players' relative strengths played an important role in chess popularity. World Chess Federation (FIDE) and many national federations, like United States Chess Federation (USCF),

[1]Universidade Federal de Alfenas - UNIFAL-MG, CEP: 37048-395, Varginha, MG, Brasil. E-mail: *danilo.pires@unifal-mg.edu.br*
[2]Universidade Federal de Lavras - UFLA, Departamento de Estatística, Caixa Posta 3037, CEP: 37209-009, Lavras, MG, Brasil. E-mail: *jssbueno@ufla.br*

keep track of players ratings resulting in very important information not only for fair pairing systems and tournament scheduling but also for marketing and sponsorship purposes.

The system used by FIDE was developed by the Hungarian-American physicist Arpad Elo, and it assumes that ratings of players are random variables and expected result is a function of rating differences. Elo has developed updating formulae using Gaussian distribution with a convenient standard deviation (and logistic approximation). Chess community follow official ratings and chess related professionals use them to describe past performance and to make predictions. From a statistical viewpoint, Elo's system is a particular case of preference models for paired comparisons (BRADLEY and TERRY,1952). One of the main generalizations of those models are contemplating ties probabilities as in (DAVIDSON, 1969) with an extra "tie propensity" parameter.

It is very easy to include in those models a general or player specific parameter to describe the "home advantage" of the player conducting white pieces. It is also possible to tune rating variances or player activity by using alternative formulae for the expected result. However, most of the comparisons for real data sets using modifications of ELO system are derived from practical realizations or computing convenience (GLICKMAN and JONES, 1999; SISMANIS, 2010; KAGGLE, 2020).

Criticism of FIDE ratings are based on well documented effects, like the larger tendency to draws in elite chess players, the advantage of playing white and the variability in players activity. This has been adressed by many authors and some very effective alternative systems have been worked out. "Glicko" system deal with heteregeneous variances on rating parameters in a Bradley-Terry type of model (GLICKMAN, 1999). A fully Bayesian approach using concepts from Glicko system is implemented in "TrueSkill" system by Microsoft Research (HERBRICH et al. 2,007). Using more frequent evaluation periods and tuning factors for beginners and experts has also been used and in fact have changed FIDE ratings calculations (FIDE, 2020a; USCF, 2020).

In what follows, we analysed actual data from elite chess players. Our objective is to find parametric statistical models that could subsidize proposals for changes in the FIDE system. Options considered are based on fitting modifications of Bradley-Terry (1952) and Davidson (1969) preference models for such unbalanced source of data. Comparison with "true" dynamic models in the literature was not considered in this context since it would be unpractical as they depend on a wider choice of prior distributions and tuning parameters as frequency of updates and shrinkage factors. Our choice instead was to emulate dynamic rating systems to make a fair comparison of different proposals. We carried out Bayesian inference using weighed likelihood functions and proper prior distributions that where elicited, and then turned vague using arbitrary higher variance.

2 Material and methods

2.1 Data

From an initial set of 28,042 official games in which both players had FIDE rating $> 2,500$ (Grandmaster level), played from January 2010 to November 2012, we selected the ones from players that would play Grand Prix series 2013 (46 players). This resulted in a *Training set* with 6,807 games that will be used to fit all considered models.

Testing sets used data from 2013 (Grand Prix, Candidates Tournament, World Cup and other major tournaments) to compare prediction abilities of the models. This was done in two different ways:

a) Using games in which both players had their rating parameters previously estimated in the training set, resulting in 411 games played by 36 players (*Testing set A*);

b) Using games in which at least one of the players had their rating parameters estimated in the training set, resulting in 732 games. For those games, 37 players had rating parameters estimated in the training set and for the remaining 51 we used current FIDE ratings (*Testing set B*).

Observable data included game date, players identification conducting white and black pieces (tournament design) and game results recorded as 0, 0.5 or 1, respectively, for a defeat, draw or win for white pieces. Game dates were used in time-dependent likelihood functions that emulate rating dynamics.

Table 3 in supplementary material has names, country of origin, and FIDE ratings (as in December 2012) of all considered elite players. A complete description (full games, biography, etc) can be found in (CHESSBASE, 2020; CHESSGAMES, 2020; CHESSRESULTS, 2020; FIDE, 2020b).

2.2 Rating models

ELO system, the proposal from BRADLEY and TERRY (1952) and derived models (hereinafter called **BT**) have a direct relation. So, let γ_i be the rating parameter in **BT** and R_i its translation in the ELO system scale. Let $y_{ij} = 1$ be the observed result for a win of player i over player j, $y_{ij} = 0.5$ for a draw and $y_{ij} = 0$ for a loss of player i. ELO (1978) formulated his model for differences of player's ratings, each normally distributed with standard deviation 200. The author then proceeded approximating the expectancy of player i score against player j in the Gaussian distribution with mean zero and standard deviation $\sqrt{2 \times 200^2} \approx 282.8$ by the inverse logistic function using normalizing factor 400 as follows:

$$E[y = 1|R_i, R_j] = \frac{1}{1 + 10^{\frac{D_{ij}}{400}}}. \quad (1)$$

In which $D_{ij} = Rj - Ri$. Note that original model has no draw probability and this result is counted as half win for each player. Thus,

$$R_i = \frac{400}{ln10}\gamma_i. \tag{2}$$

This is the same as a preference probability for i over j in **BT**. For easy reading among chess players and arbiters, we present rating estimates corrected to FIDE-ELO's scale, like above.

Modifying the model to include a single parameter for all players that represents a common advantage of playing white will result in the following change in the linear predictor: $D^*_{ij} = R_j - R_i - \delta$. Similarly, it is possible to make a new linear predictor $D^{**}_{ij} = R_j + \delta_j - R_i - \delta_i$, allowing for different advantages of being white for each player.

Under this specification, likelihood function for **BT** is written as:

$$\mathbf{L_{BT}} = \prod_{k=1}^{n} \pi_{i_k j_k}^{y_{ijk}} (1 - \pi_{ijk})^{(1-y_{ijk})}, \tag{3}$$

where π_{ijk} is the expected result in favor of white player to win the game and y_{ijk} is the result from k^{th} game between players i (as white) versus j (as black).

The proposal from Davidson (1969) and derived models are hereinafter called **DV**. They differ from **BT** by having three preference classes, directly modeling the probability of draw. Considering it is not very direct to re-scale parameters of those models to meet FIDE-ELO ratings, we kept them in the original scale.

Below are described models with a common white advantage parameter δ. Original model only has one parameter λ related to the drawing propensity. In what follows, this model can be yielded back making $\delta = 0$. To have models with specific white advantage parameters for each player we simply replace δ by δ_i and δ_j.

$$\begin{aligned}
\pi_{ij} = P(i\ win\ j) &= \frac{e^{\gamma_i+\delta}}{e^{\gamma_j} + e^{\gamma_i+\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j+\delta}{2}}}, \\
\pi_{ij0} = Pr(i\ draw\ j) &= \frac{e^{\lambda+\frac{\gamma_i+\gamma_j+\delta}{2}}}{e^{\gamma_j} + e^{\gamma_i+\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j+\delta}{2}}}, \\
\pi_{ji} = P(i\ loose\ j) &= \frac{e^{\gamma_j}}{e^{\gamma_j} + e^{\gamma_i+\delta} + e^{\lambda+\frac{\gamma_i+\gamma_j+\delta}{2}}}.
\end{aligned} \tag{4}$$

For **DV** is possible to rewrite the likelihood function as:

$$\mathbf{L_{DV}} = \prod_{k=1}^{n} \left(\pi_{ijk}^{I_{\{y_{ijk}=1\}}} \pi_{ijk0}^{I_{\{y_{ijk}=0,5\}}} \pi_{jik}^{I_{\{y_{ijk}=0\}}} \right), \tag{5}$$

where $I_{\{.\}}$ is an indicator variable for game result.

Gaussian prior distributions were used for ratings and white advantages. Their averages were elicited, but with high variances, to allow for easy learning from data. For **BT** models, μ hyperparameter was the quantile of FIDE-ELO for the 100^{th} best player and σ was twice the one used for ELO system. This preserves a proper prior distribution with little information on ratings. For δ, we used a small but positive average as expected from chess literature, reflecting about 7% increase in the winning probability for white, with a large standard deviation to allow for negative values. This also reflected in a proper prior with little information. Specific distributions chosen for **BT** are: $R_i \sim N(\mu = 2705, \sigma = 400)$, $\delta_i \sim N(\mu = 50, \sigma = 40)$ and $\delta \sim N(\mu = 50, \sigma = 40)$.

For **DV** models those hyperparameters were just re-scaled, and λ has also a small but positive value (reflecting common knowledge that draws are more frequent in games between elite players) with large standard deviation allowing proper priors with little information. For **DV**s we chose following prior distributions: $\gamma_i \sim N(\mu = 15, \sigma = 400)$, $\delta_i \sim N(\mu = 1, \sigma = 10)$, $\delta \sim N(\mu = 1, \sigma = 10)$ and $\lambda \sim N(\mu = 1, \sigma = 5)$.

Joint posterior distributions for each model are products of Gaussian priors and Bernoulli (**BT**) or multinomial (**DV**) likelihoods. The full conditional distributions can be simplified as follows:

BT: Distribution of rating for the i^{th} player (R_{i^*}) given all other player's ratings ($\mathbf{R}_{-i^*}$) and player-specific white advantage parameters δ:

$$P(R_{i^*} \mid \mathbf{R}_{-i^*}, \delta, \mathbf{y}) \propto \mathbf{L_{BT}} \times e^{\frac{-(R_{i^*}-2705)^2}{320000}} \tag{6}$$

Distribution of the white advantage parameter for the i^{th} player δ_i given all other player's ($\delta_{\mathbf{i}}$) and all player's ratings ($\mathbf{R}$):

$$P(\delta_{i^*} \mid \mathbf{R}, \delta_{-i^*}, \mathbf{y}) \propto \mathbf{L_{BT}} \times e^{\frac{-(\delta_{\mathbf{i}^*}-50)^2}{3200}} \tag{7}$$

In the same fashion, we present full conditional distributions for **DV** as follows:

DV: Distribution of rating for the i^{th} player (γ_{i^*}) given all other player's ratings ($\mathbf{gamma}_{-i^*}$), player-specific white advantage parameters δ and drawing propensity parameter (λ):

$$P(\gamma_i, \mid \gamma_{-i}, \delta, \lambda, \mathbf{y}) \propto \mathbf{L_{DV}} \times e^{\frac{-(\gamma_i-15)^2}{32000}} \tag{8}$$

Distribution of white advantage parameters δ_i for the i^{th} given all other player's (δ_{-i^*}), player's ratings (γ) and drawing propensity parameter (λ):

$$P(\delta_{i^*} \mid \gamma, \delta_{-i^*}, \lambda, \mathbf{y}) \propto \mathbf{L_{DV}} \times e^{\frac{-(\delta_i-1)^2}{200}} \tag{9}$$

Distribution of drawing propensity parameter (λ) given player's ratings (**gamma**) and white advantage parameters δ:

$$P(\lambda \mid \gamma, \delta, \mathbf{y}) \propto \mathbf{L_{DV}} \times e^{\frac{-(\lambda-1)^2}{50}} \tag{10}$$

2.3 Sampling from the joint posterior distribution

All inference on the descriptive properties of the models was carried out on samples from joint marginal distribution for all parameters given training data. Samples from predictive distributions for each model were also drawn. This made possible to evaluate model performance in the *Testing sets*. Markov Chain Monte-Carlo methods were used to sample the joint posterior distributions. Sampling from posterior distributions was done using Metropolis-Hastings methods (HASTINGS,1970) with adaptive candidate generating functions embedded within Gibbs Sampling algorithm (GILKS et al., 1995).

Normal distributions were used to generate candidate points. Hyperparameters from those distributions were updated after a window of $1,000$ samples. In what follows, we present marginal summaries from the posterior distributions for each model and it's parameters.

For each parameter 3 parallel chains with $130,000$ iterations were drawn. We burnt the first $50,000$ samples and later used a 20 iterations jump, yielding final sample size of $4,000(\times 3)$.

Algorithms were implemented in R (R CORE TEAM R, 2020) and sampling diagnostics were done using (RAFTERY and LEWIS, 1992), (GELMAN and RUBIN, 1992) and (BROOKS and GELMAN, 1998) criteria implemented on *coda* package (PLUMMER et al. 2006).

2.4 Weighed Likelihood to emulate dynamic models

A time related variable ω was created according to Sismanis (2010) to weigh the likelihood function in a way that old games has less importance than new games.

$$\omega_t = \frac{1}{\sqrt{\frac{t-t_{min}}{t_{max}-t_{min}}}}\ , \tag{11}$$

in which t_{min} and t_{max} are, respectively, the time of a game being analyzed and the time considered as reference to make inference in the *Training set*.

Using ω in the likelihood function makes for a proxy of true dynamic models. Other ways to implement dynamic models would require to carry out the estimation computing posterior averages (or modes) periodically or making a fully parameterized model in which ratings have a parameter for each time. Those models would be too much of a computational burden to the purpose of this paper

whose main objective is to elucidate the effect of having such a correction on rating models. So we just have implemented time weighing leaving actual dynamic models for future research. However, we will refer to models with ω weighing the likelihood as "dynamic models", in contrast to "static models" with $\omega = 1$. Some results are presented for all 12 models (6 static and 6 dynamic), while others are just presented for the best version of **BT** or **DV**.

2.5 Decision on best models

Decision on best models, both from description and for prediction was based on direct estimates of AIC (AKAIKE,1974) or its Monte Carlo approximation AICM (RAFTERY and NEWTON,2007). For prediction models in each scenario we used two different ideas on how to handle incomplete data:

a) The joint predictive distributions for possible game result from *Testing sets* A and B were worked out. This yielded probability estimates of $p(y^*|y, \theta)$, in which y^* is the new result actually observed in testing sets;

b) Based on $p(y^*|y, \theta)$ we evaluated information or distance criteria to realized y^* in each given model. Those results are numerical approximations of predictive AIC, AICM and DeFinetti distance measures, in each case.

For predictions using *Testing set B*, for each player that had not been monitored, we used ELO rating as given by FIDE in the game day (eventually rescaled to contemplate model specification in **DV**). As for δ parameter, the following proposals were used:

i $\delta_i = 0$, representing no advantage of conducting white;

ii $\delta_i = \mu_\delta$, or the posterior marginal average for the advantage of playing as white, found for the model with a single δ;

iii $\delta_i = -\delta_j$, assuming the estimate for players j could be considered as reverse effect in his opponent.

Each result was compared to a correspondent reference model using educated guesses (or naive estimation of probabilities). Assumptions for each of those references are:

I a win, a draw or a loss are equiprobable, meaning basic $50\% - 50\%$ to **BT** or $\frac{1}{3}, \frac{1}{3}$ and $\frac{1}{3}$ to **DV**;

II summaries of historic results for wins, draws and losses from white players' perspective. This "proportional" model can be used to check **DV**.

III the same as previous case, but considering a draw as half a loose and half a win to check **BT**, as it is basically a rescaled ELO.

As a special case to check **BT**, we used probabilities derived from FIDE ratings as given in the game day. This represents a true dynamic model that is expected to outperform its static counterpart. To make fair predictions, however, we used FIDE ratings from December 2012 for *Testing sets A* and *B*.

3 Results and discussion

Samples from joint posterior distribution had good statistical properties and could be treated as independent. Examples are depict in trace plots from Figures 1 (Michael Adams, **BT**) and 2 (Shakhriar Mamediarov, **DV**). That made simpler to evaluate all *post hoc* statistics or depict plots from marginal posterior distributions. Prior information was not relevant to inference as can be seen in the densities presented in Figures 3 and 4 for respective players and models.

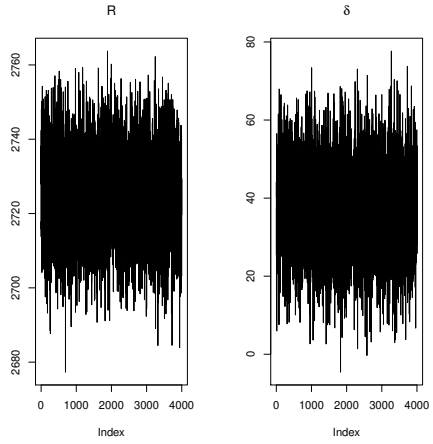

Figure 1 - Trace plot for samples in posterior distribution for Michael Adams' rating and a comon white advantage parameter. **BT** model with weighed likelihood (ω).

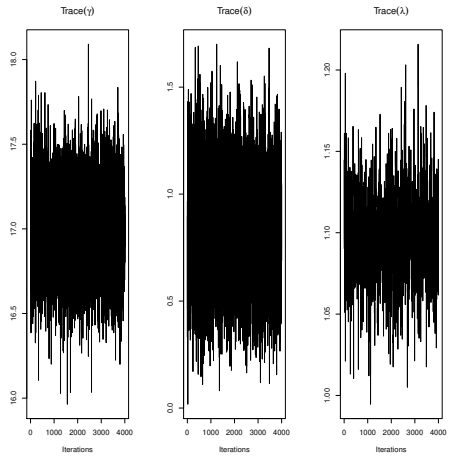

Figure 2 - Trace plot for samples in posterior distribution for Shakhriar Mamediarov's ratings, with common white advantage and drawing propensity parameter. **DV** model with weighed likelihood (ω).

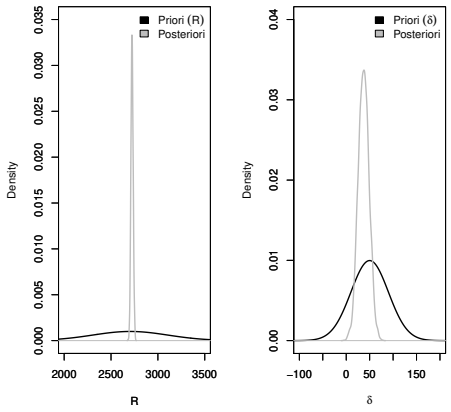

Figure 3 - Posterior density for Michael Adams' rating (grey), and respective prior density (black). **BT** model with a common white advantage parameter (δ) and weighed likelihood (ω).

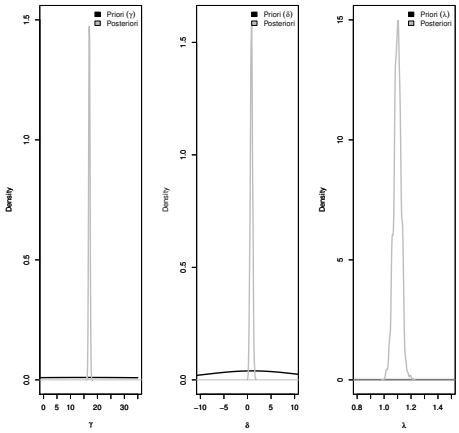

Figure 4 - Posterior density for Shakhriar Mamediarov's parameters (grey), and respective prior densities (black). **DV** model with a common white advantage parameter (δ) and weighed likelihood (ω).

3.1 Posterior analysis

In Table 1 we present estimates of AIC from different models fitted to *Training set* and evaluated in *Testing sets*. Such estimates can be used for model comparison by themselves or compared to static references and FIDE-ELO official ratings used as dynamic reference.

As we can see all the proposed models outperform equivalent reference model. Regarding **BT**, the best choice for both static and dynamic models was to use a single parameter to the advantage of playing white. Posterior average for the marginal distribution for this parameter in static model was $\hat{\delta} = 41.7$ with a 95% HPD given by [33.5 ; 50.6]. In the dynamic model it was estimated as $\hat{\delta} = 43.7$, with HPD given by [37.1 ; 50.1]. This is roughly equivalent to say that a player has a 41.7 increase in its ELO rating if it is assigned to play as white (using static model), or a 43.7 increase in the dynamic model.

As an example of the static model use, in a game between players A, $R_i = 2,686$ and B, $R_j = 2,715$, first player would have a 6% increase in its expected winning probability by playing white, or equivalently, a $2,727.7$ strength. The same example in the dynamic model would result in a $2,729.7$ strength and 6.3% increase in the winning chance for A.

In Figure 5 we have shown that ELO ratings are equivalent to **BT** models and underestimate probabilities of white victory. **BT** models with a common white advantage parameter has shown better fit to actual data. ELO ratings specially underestimate white advantage with high rating differences ($\Delta_R < -200$ or $\Delta_R > 200$).

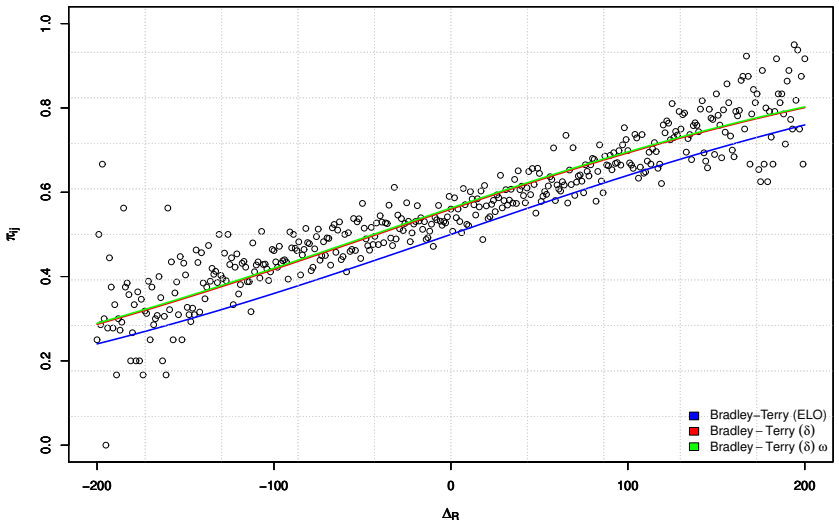

Figure 5 - Estimated probability of white's victory $\hat{\pi}_{ij}$ as a function of rating differences Δ_R. Models: basic **BT** (ELO) (blue line), **BT** with a single $\hat{\delta} = 41.7$ (red line) and **BT** with a single $\hat{\delta} = 43.7$ and weighed likelihood (green line)

Table 1 - AIC estimates (the smaller the better) for proposed and reference models. Static models has $\omega_t = 1$ and dynamic models has unknown ω_t. In boldface are highlighted smaller values (best models) for each section.

Models		Parameters	AIC
Bradley-Terry	$\omega_t = 1$	$\theta = R_i$	8936.4
		$\theta = R_i, \delta$	**8847.2**
		$\theta = R_i, \delta_i$	8878.6
		Equiprobable	9438.5
		Proportional with no draws	9474.8
		Proportional	11483.2
		FIDE-ELO	**8947.4**
	ω_t	$\theta = R_i$	15841.8
		$\theta = R_i, \delta$	**8847.2**
		$\theta = R_i, \delta_i$	15644.9
		Equiprobable	16788.8
		Proportional with no draws	16835.3
		Proportional	20458.2
		ELO	**15901.6**
Davidson	$\omega_t = 1$	$\theta = \gamma_i, \lambda$	12686.9
		$\theta = \gamma_i, \delta, \lambda$	12477.3
		$\theta = \gamma_i, \delta_i, \lambda$	**12425.2**
		Equiprobable	14956.5
		Proportional	**13573.1**
	ω_t	$\theta = \gamma_i, \lambda$	23510.3
		$\theta = \gamma_i, \delta, \lambda$	21961.1
		$\theta = \gamma_i, \delta_i, \lambda$	**21887.3**
		Equiprobable	26610.4
		Proportional	**24129.0**

All models derived from Davidson basic version were better than respective reference models. Best model has both players white advantage parameters and a single drawing parameter, that was estimated as $\lambda = 1,098$ with $HPD_{95\%} : [1,04 ; 1,15]$ for the static model and as $\lambda = 1,106$ with $HPD_{95\%} : [1,06 ; 1,14]$ for the model with weighed likelihood.

Those estimates are indicative of a prevalent drawing tendency in games played at this level. Comparing these with rating estimates that one can find in Figure 6, estimated by the best models (static and dynamic versions) is a good indicative of the extent those players are prone to draw.

For instance, take the probabilities for expected outcomes of current world champion Magnus Carlsen (with highest rating $\gamma \approx 18$ or $ELO \approx 3120$) playing his predecessor Ruslan Ponomariov (with rating estimated as $\gamma \approx 16$ or $ELO \approx 2780$). Ignoring white advantage, the expected result would be approximately 70%. But a win for white would be just as likely as a draw, with around 46.8% probability.

Consider now two much weaker players with the same rating difference, lets say, $\gamma_1 = 12$ ($ELO \approx 2084$) and $\gamma_2 = 10$ ($ELO \approx 1787$). Even if rating differences in ELO scale are smaller, due to the proportionality factor, winning chances for white are higher. In this case, expected result would be about 78.8% with a winning probability of 66.5% for the first player. This is in strong agreement of what is observed in practice.

White advantage parameter estimates by the two best versions of **DV** are depicted in Figure 7. In here, we found a considerable disagreement between both models. For the static model, estimates seems to imply that there is a small group of players that considerably benefit from playing white. For most of the other players, credibility intervals include ($\delta = 0$). It is worth noting that just Ponomariov and Andreikin apparently do not perform better than expected as white. It is possible that due to weaker opposition their results can be biased, but the result for Mamediarov, Bacrot, Caruana, Shirov and Kramnik are very consistent with practical observation. As a matter of fact, Topalov's estimate having such a large credibility interval (including $\delta = 0$ as likely) is slightly unexpected, but may reflect his diminished activity in the period.

The estimates from dynamic model, on the other hand, has shown a larger group of players that benefit clearly from playing white, being Ponomariov the lone exception and Topalov the most prone to win as white. Again, based on reports from numerous tournaments and psychological attributes fellow players assign to ex-world champions, it is a likely result.

Posterior distributions were far off the prior averages. This is clear sign of likelihood dominance over prior information,

3.2 Predictive analysis

Testing set A

In this section we evaluated prediction for new games involving players that have parameters estimated in *Training set*. Table 2 brings AIC estimates for each

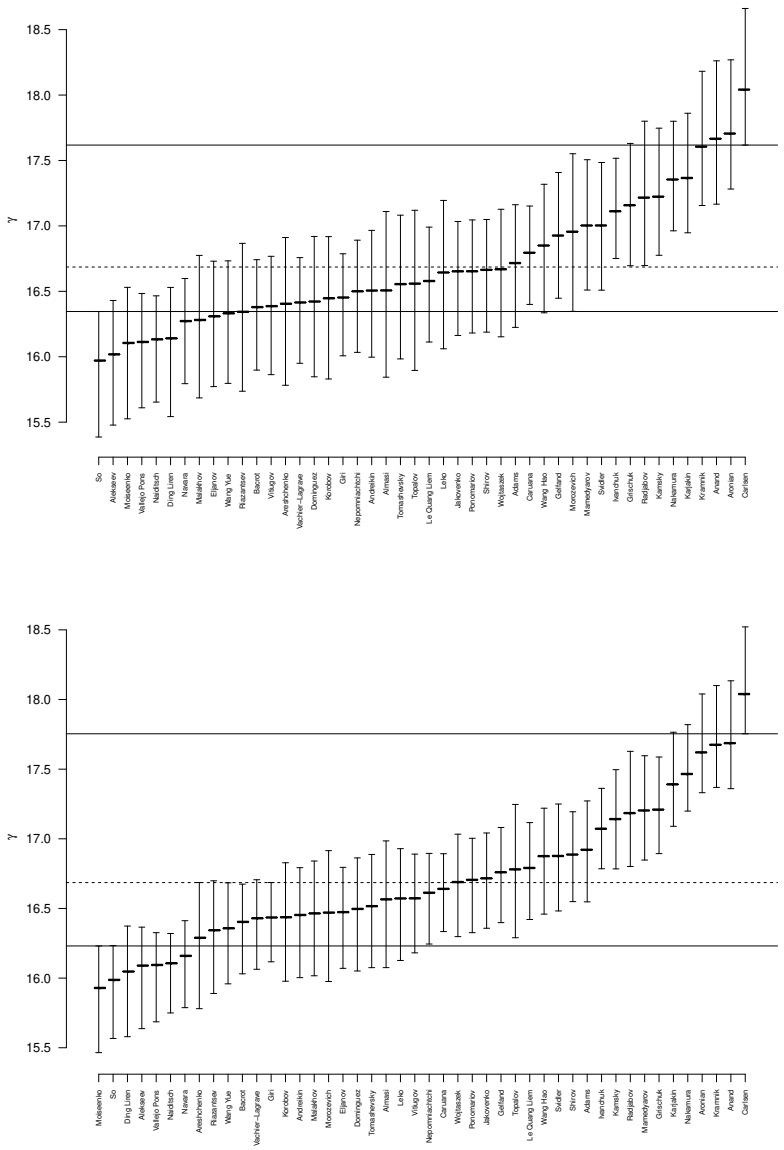

Figure 6 - Ratings in best **DV** models (γ_i, δ_i and λ) showing higher power for dynamic version. Posterior averages and 95% HPD credibility intervals for player's rating parameters γ_i. Horizontal line indicates average of the group. On top is the static and on bottom the dynamic version of the model.

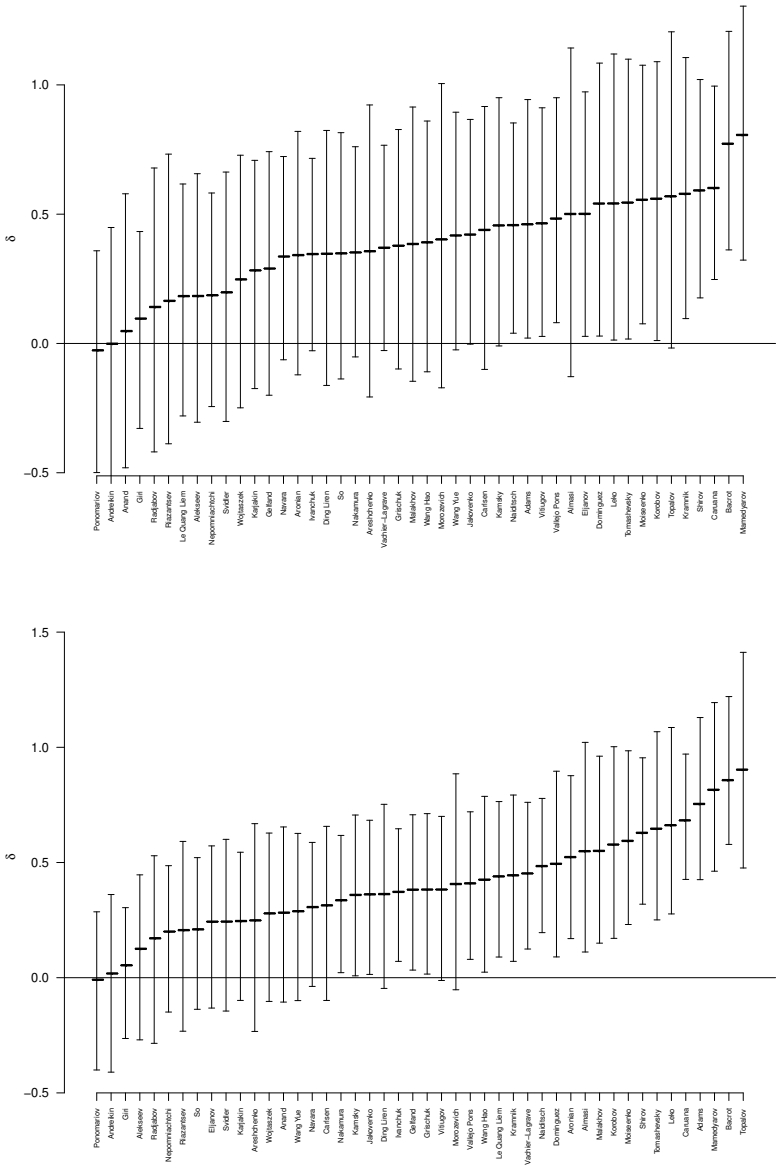

Figure 7 - White advantage in best **DV** models (γ_i, δ_i and λ) showing higher power for dynamic version. Posterior averages and 95% HPD credibility intervals for player's white advantage parameters δ_i. Horizontal line indicates average of the group. On top is the static and on bottom the dynamic version of the model.

model applied to both *Testing sets*. For *testing set A*, In all considered cases **BT** dynamic versions outperform the respective static version. One **BT** and two **DV** models beat their respective references. In both cases, best model was the dynamic version with a time-weighed likelihood, and a common white advantage parameter.

Note that ELO ratings are truly dynamic in its implementation and outperform other references. This result agrees with recent findings in Delloite Challenges using **BT** dynamic models (Sismanis, 2010). This is a strong hint that true dynamic models (ratings evaluated continuously) would be better proxies to chess strength than current system.

Testing set B

In *Testing set B* there are games in which at least one of the players have parameters estimated in *Training set* ("first player"). To evaluate predictions it was necessary to input values for "new players" that where not in the *Training set*. We used FIDE-ELO ratings declared for the game day and three options for individual δ_j parameters, namely:

- $\delta_j = 0$: using no advantage for the new player;
- $\delta_j = -\delta_i$: using minus the advantage estimated for the "first player";
- $\delta_j = \mu_\delta$: using the average of the advantages estimated for all the players.

Using both **BT** and **DV** models, the best prediction comes from the ones with a common parameter for white advantage. Those are also the only models that outperform all the references in all the criteria. This is an indication that if, for any reason, individual parameters delta are badly estimated (or estimates are unavailable) we should use a simpler model with a common parameter δ for better prediction.

The good features of the ELO dynamic strategy can be improved easily with a single parameter for white advantage. This could be evaluated for different basis and purposes, for instance, based on games played in a given year, tournament or league.

The use of white advantage parameters enhances the fit in all versions. For past performance, using **BT** we should keep a single parameter (δ) for white advantage, but using **DV**, individual parameters for each player (δ_i) is the best choice.

Fully parameterized **DV** may enlighten an interaction of individual white advantage and the well known effect of increasing drawing probability (λ) with the average strength of the players. Those models are more difficult to fit, but could help to better describe past performance and also subsidize more accurate predictive systems.

We successfully implemented Bayesian inference for all considered models using R (R CORE TEAM, 2020). Efficiency of the algorithms was not of great concern as we are not advocating direct use of the models to a new rating system. However,

Table 2 - AIC estimates for proposed and reference models in *Testing sets A* and *B*. We used boldface to disclose smaller values in each model type. Static models has $\omega_t = 1$ and dynamic versions varying ω_t.

Model	Parameters			Testing Sets	
				A	B
BT			$\theta = R_i$	576.99	1008.11
			$\theta = R_i, \delta$	570.63	998.69
	$\omega_t = 1$	$\delta_i = 0$	$\theta = R_i, \delta_i$	1005.90	
		$\delta_i = \mu_{\delta_s}$	$\theta = R_i, \delta_i$	576.33	1009.61
		$\delta_i = -\delta_f$	$\theta = R_i, \delta_i$		1028.30
	ω_t		$\theta = R_i$	573.73	1005.06
			$\theta = R_i, \delta$	**567.30**	**995.10**
		$\delta_i = 0$	$\theta = R_i, \delta_i$		997.16
		$\delta_i = \mu_{\delta_s}$	$\theta = R_i, \delta_i$	571.76	1000.46
		$\delta_i = -\delta_j$	$\theta = R_i, \delta_i$		1012.79
			Equiprobable	573.76	1016.76
			Proportional, no ties	577.63	1015.09
			Proportional	712.37	1179.19
			ELO	570.46	1031.60
DV			$\theta = \gamma_i, \lambda$	810.64	1542.32
			$\theta = \gamma_i, \delta, \lambda$	795.42	**1527.77**
	$\omega_t = 1$	$\delta_i = 0$	$\theta = \gamma_i, \delta_i, \lambda$		1592.32
		$\delta_i = \mu_{\delta_s}$	$\theta = \gamma_i, \delta_i, \lambda$	809.39	1598.47
		$\delta_i = -\delta_j$	$\theta = \gamma_i, \delta_i, \lambda$		1637.11
	ω_t		$\theta = \gamma_i, \lambda$	808.17	1548.47
			$\theta = \gamma_i, \delta, \lambda$	**792.83**	1536.01
		$\delta_i = 0$	$\theta = \gamma_i, \delta_i, \lambda$		1583.53
		$\delta_i = \mu_{\delta_s}$	$\theta = \gamma_i, \delta_i, \lambda$	803.94	1590.32
		$\delta_i = -\delta_j$	$\theta = \gamma_i, \delta_i, \lambda$		1671.00
			Equiprobable	907.27	1621.36
			Proportional	793.66	1537.39

the analysis presented in here reinforces the idea that FIDE-ELO system should be revised. Dynamic models using parameters for white advantage and for drawing probabilities would be a good basis for an alternative system.

4 Conclusion

Comparing past results and predictions, we could show that the use of weighed likelihood functions is helpful to verify the advantages of dynamic models.

A new system with a single white advantage parameter (δ) could be used to improve ELO. Both **BT** or **DV** could be used as a basis for such system. The former is closer to current FIDE ratings, but the later has best predictive potential.

Acknowledgments

We would like to thank reviewers and editors for their comments and suggestions. The authors would also like to thank the doctoral scholarship granted by CAPES to the first author, and the research aid granted by FAPEMIG to the second author.

PIRES, D. M.; BUENO FILHO, J. S. S. Podemos melhorar os ratings ELO? Um estudo de caso com enxadristas de elite. *Rev. Bras. Biom.,* Lavras, v.38, n.4, p.483-505, 2020.

- *RESUMO: Originalmente planejados para descrever o desempenho pregresso, os sistemas de ratings do xadrez agora são amplamente usados para refletir a força relativa dos jogadores, com muitos aspectos importantes na programação de torneios, publicidade e premiações. O sistema ELO foi adotado oficialmente pela Federação Mundial de Xadrez (FIDE). Implementamos a análise bayesiana de resultados de jogos da elite do xadrez mundial para ajustar modelos estatísticos paramétricos que poderiam subsidiar propostas de melhorias no sistema. Embora a maioria das opções consideradas não seja nova, mas baseadas em modelos bem conhecidos de preferência, o uso de verossimilhança ponderada para emular sistemas dinâmicos na maneira que implementamos a inferência bayesiana é novo. Nós comparamos a capacidade descritiva usando critérios de informação baseados na verossimilhança marginal. O critério de informação de Akaike foi utilizado para comparar as predições. Muitas das opções consideradas melhoram o sistema ELO e há fortes evidências de que modelos dinâmicos, considerando tanto a vantagem das brancas quanto a propensão a empates podem resultar em sistemas mais acurados.*

- *PALAVRAS-CHAVE: Inferência bayesiana; avaliação de desempenho; modelos de preferência; esportes.*

References

AKAIKE, H. A new look at the statistical model identification, *Transactions on Automatic Control*, v.19, no. 6, pp. 716-723, 1974.

BRADLEY, R. A.; TERRY, M. E. Rank analysis of incomplete blocks designs :i. the method of paired comparisons. *Biometrika* v.39, n.3, p.324-345,1952.

BROOKS, S. P.; GELMAN, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, v. 7, n. 4, p. 434-455,1998.

CHESSBASE-Chessbase news. Accessed: 2020-06-24. URL:*https://en.chessbase.com/post/fide-grand-prix-2013-in-lisbon-madrid-berlin-paris*, 2020.

CHESSGAMES-Chess masters final 4th bilbao. Accessed: 2020-06-24. URL: *https://www.chessgames.com/perl/chess.pl?tid=79025*,2020

CHESSRESULTS-Chessresults the international chess-tournaments results server'. Accessed: 2020-06-24. URL: *http://chess-results.com*,2020.

DAVIDSON, R. R. (1969), On extending the bradley-terry model to accommodate ties in paired comparison experiments'. *Journal of American Statistical Association* v.65, n.37, p.317-328, 1969.

ELO, A. E. The rating of chess players past and present. New York: Arco Publishing, 1978.

FIDE - Federation internationale des echecs. Accessed: 2020-06-24. URL:*http://www.fide.com*, 2020a.

FIDE - Ratings Fide. Accessed: 2020-06-24. URL:*http://ratings.fide.com*, 2020b.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* v.7, p.457-511, 1992.

GILKS, W. R.; BEST, N. G.; TAN, K. K. C. "Adaptive Rejection Metropolis Sampling within Gibbs Sampling". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* v.44, n.4, p.455–472,1995.

GLICKMAN, M. E. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, v. 48, p.377-394, 1999.

GLICKMAN, M. E.; JONES, A. C. Rating the chess ratings system. *Chance* v.12, p. 21-28, 1999

HASTINGS, W. K. Monte Carlo Sampling Methods using Markov chains and their applications. *Biometrika*, v.57, n.1, p. 97-109, 1970.

HERBRICH, R.; MINKA, T.; GRAEPEL, T. Trueskill(tm):A bayesian skill rating system.*MIT Press*, p. 569-576, 2007.

KAGGLE. Chess ratings - Elo versus the Rest of the World. Accessed: 2020-06-24. URL: *https://www.kaggle.com/c/chess/data*, 2020.

MARTINHO, L.; YANG, H.; LUENGO, D.; KANNIAINEN, J.; CORANDER, J. A fast universal self-tuned sampler within Gibbs sampling.*Special Issue in Honour of William J. (Bill) Fitzgerald*. v.47, p.68–83, 2015.

PLUMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, v.6,n.1, p. 7-11.

RAFTERY, A. E.; LEWIS, S. M. Comment: one long run with diagnostics: implemetation strategies for markov chain monte carlo. *Statical Science*, v. 7, n.4, p.493-497, 1992.

RAFTERY, A. E.; NEWTON, M. A. Estimating the integrated likehood via posterior simulation using the harmonic mean identity. *Bayesian Statistics* v.8, p.1-45,2007.

R CORE TEAM . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/, 2020.

SISMANIS, Y. How I won the "Chess Ratings-Elo vs the Rest of the World" competition. *CoRR, abs/1012.4571*, 2010.

USCF- United states chess federation. Accessed: 2020-06-24.
URL: *http://www.uschess.org/content/blogsection/14/195/*,2020.

Received on 13.12.2019.

Approved after revised on 29.07.2020.

SUPPLEMENTARY MATERIAL

Web page: The full description one of the models can be found in https://danilomachadopires.github.io/index

R-code: R-code for all models is also provided in this web page

Data sets: Four data sets will be provided in the web page as R objects (.rda files)

Table 3 - Players names, country of origin and FIDE ratings as in December 2012. Last columns indicate which dataset has the player's games: *TR*=*Training set* (6,807 games), *A*=*Testing set A* (411 games) and *B*=*Testing set B* (732 games).

	Name	Country	Rating	*TR*	*A*	*B*		Name	Country	Rating	*TR*	*A*	*B*
1	Carlsen, Magnus	NOR	2848	x	x	x	49	Almasi, Zoltan	HUN	2689	x		
2	Aronian, Levon	ARM	2815	x	x	x	50	Grachev, Boris	RUS	2688			x
3	Kramnik, Vladimir	RUS	2795	x	x	x	51	Movsesian, Sergei	ARM	2688			x
4	Radjabov, Teimour	AZE	2793	x	x	x	52	Rublevsky, Sergei	RUS	2688			x
5	Caruana, Fabiano	ITA	2782	x	x	x	53	Eljanov, Pavel	UKR	2687	x		
6	Anand, Viswanathan	IND	2775	x	x	x	54	So, Wesley	PHI	2682	x		
7	Karjakin, Sergey	RUS	2775	x	x	x	55	Bologan, Viktor	MDA	2681			x
8	Topalov, Veselin	BUL	2771	x	x	x	56	Fridman, Daniel	GER	2667			x
9	Ivanchuk, Vassily	UKR	2766	x	x	x	57	Sargissian, Gabrie	ARM	2666			x
10	Grischuk, Alexander	RUS	2764	x	x	x	58	Potkin, Vladimir	RUS	2665			x
11	Mamedyarov, Shakhriyar	AZE	2764	x	x	x	59	Granda Zuniga, Julio E	PER	2664			x
12	Kamsky, Gata	USA	2762	x	x	x	60	Georgiev, Kiril	BUL	2660			x
13	Nakamura, Hikaru	USA	2760	x	x	x	61	Onischuk, Alexander	USA	2660			x
14	Gelfand, Boris	ISR	2751	x	x	x	62	Safarli, Eltaj	AZE	2660			x
15	Morozevich, Alexander	RUS	2748	x	x	x	63	Khenkin, Igor	GER	2659			x
16	Svidler, Peter	RUS	2747	x	x	x	64	Khairullin, Ildar	RUS	2658			x
17	Jakovenko, Dmitry	RUS	2741	x	x	x	65	Khismatullin, Denis	RUS	2658			x
18	Wang, Hao	CHN	2737	x	x	x	66	Zvjaginsev, Vadim	RUS	2658			x
19	Wojtaszek, Radoslaw	POL	2734	x			67	Kurnosov, Igor	RUS	2657			x
20	Dominguez Perez, Leinier	CUB	2734	x	x	x	68	Dreev, Aleksey	RUS	2654			x
21	Leko, Peter	HUN	2732	x	x	x	69	Kobalia, Mikhail	RUS	2652			x
22	Ponomariov, Ruslan	UKR	2732	x	x	x	70	Smirin, Ilia	ISR	2652			x
23	Tomashevsky, Evgeny	RUS	2725	x	x	x	71	Tkachiev, Vladislav	FRA	2649			x
24	Andreikin, Dmitry	RUS	2723	x	x	x	72	Dubov, Daniil	RUS	2638			x
25	Areshchenko, Alexander	UKR	2720	x			73	Mamedov, Rauf	AZE	2637			x
26	Giri, Anish	NED	2720	x	x	x	74	Najer, Evgeniy	RUS	2633			x
27	Vachier-Lagrave, Maxime	FRA	2711	x	x	x	75	Popov, Ivan	RUS	2632			x
28	Adams, Michael	ENG	2710	x	x	x	76	Guseinov, Gadir	AZE	2631			x
29	Navara, David	CZE	2710	x			77	Nguyen, Ngoc Truong Son	VIE	2625			x
30	Moiseenko, Alexander	UKR	2710	x	x	x	78	Meier, Georg	GER	2610			x
31	Malakhov, Vladimir	RUS	2709	x			79	Ponkratov, Pavel	RUS	2605			x
32	Cheparinov, Ivan	BUL	2709			x	80	Mecking, Henrique	BRA	2604			x
33	Shirov, Alexei	LAT	2708	x			81	Rakhmanov, Aleksandr	RUS	2602			x
34	Naiditsch, Arkadij	GER	2708	x		x	82	Frolyanov, Dmitry	RUS	2570			x
35	Nepomniachtchi, Ian	RUS	2707	x	x	x	83	Ghaem Maghami, Ehsan	IRI	2554			x
36	Le, Quang Liem	VIE	2705	x	x	x	84	Vasquez Schroeder, Rodrigo	CHI	2542			x
37	Riazantsev, Alexander	RUS	2705	x	x	x	85	Salem, A.R. Saleh	UAE	2531			x
38	Akopian, Vladimir	ARM	2704			x	86	Artemiev, Vladislav	RUS	2524			x
39	Bacrot, Etienne	FRA	2703	x	x	x	87	Gundavaa, Bayarsaikhan	MGL	2516			x
40	Ding, Liren	CHN	2702	x	x	x	88	Pridorozhni, Aleksei	RUS	2512			x
41	Korobov, Anton	UKR	2702	x	x	x	89	Rodriguez Vila, Andres	URU	2508			x
42	Fressinet, Laurent	FRA	2700			x	90	AL-Sayed, Mohammed	QAT	2507			x
43	Kasimdzhanov, Rustam	UZB	2696			x	91	Gordievsky, Dmitry	RUS	2474			x
44	Wang, Yue	CHN	2696	x	x	x	92	Potapov, Pavel	RUS	2460			x
45	Vallejo Pons, Francisco	ESP	2694	x	x	x	93	Cherniaev, Alexander	RUS	2447			x
46	Vitiugov, Nikita	RUS	2694	x	x	x	94	Nadanian, Ashot	ARM	2428			x
47	Inarkiev, Ernesto	RUS	2693			x	95	Pasiev, Rakhim	RUS	2384			x
48	Alekseev, Evgeny	RUS	2691	x			96	Sibriaev, Aleksandr	RUS	2239			x

Table 4 - Rating estimates for the best **BT** and **DV** models for each of the 46 considered players. FIDE-ELO ratings for 2012 (December), and 2014 (May). R_s and R_d are rating estimates for static and dynamic **BT** models, γ are rating estimates for static and dynamic **DV** models, δ are white advantages estimates for static and dynamic **DV** models, γ(ELO) e δ(ELO) are estimates transformed to ELO's scale.

	Names	ELO ratings		BT		DV $\omega_t = 1$				DV ω_t			
		2012	2014	R_s	R_d	γ	$\gamma(ELO)$	δ	$\delta(ELO)$	γ	$\gamma(ELO)$	δ	$\delta(ELO)$
1	Adams, Michael	2710	2750	2729.1	2741.5	16.69	2899.3	0.47	80.88	16.91	2938.3	0.77	134.17
2	Alekseev, Evgeny	2691	2673	2680.3	2688.2	15.94	2768.4	0.17	28.83	16.02	2782.3	0.09	16.21
3	Almasi, Zoltan	2689	2693	2714.8	2720.8	16.46	2860.1	0.51	88.36	16.53	2871.5	0.55	95.71
4	Anand, Viswanathan	2775	2785	2818.2	2820.3	17.72	3077.6	0.02	3.39	17.74	3081.9	0.26	45.64
5	Andreikin, Dmitry	2723	2722	2716	2711.6	16.46	2859.9	-0.03	-5.81	16.41	2850.5	-0.02	-3.89
6	Areshchenko, Alexander	2720	2701	2701.5	2692	16.35	2841	0.35	61.33	16.23	2819.8	0.23	39.36
7	Aronian, Levon	2815	2815	2816.8	2812.5	17.76	3085.1	0.34	58.51	17.67	3069.5	0.52	90.89
8	Bacrot, Etienne	2703	2721	2707.5	2706.4	16.33	2836.2	0.8	139.31	16.36	2841.2	0.88	153.55
9	Carlsen, Magnus	2848	2882	2844.8	2846	18.12	3148.1	0.44	76.79	18.12	3148	0.3	51.6
10	Caruana, Fabiano	2782	2783	2740.4	2726.9	16.78	2914.1	0.62	107.21	16.61	2885.7	0.7	120.81
11	Ding, Liren	2702	2714	2686.7	2678.5	16.07	2791.4	0.34	59.49	15.97	2774.3	0.35	60.78
12	Dominguez Perez, Leinier	2734	2768	2720.8	2728.1	16.37	2844.2	0.55	95.93	16.46	2858.7	0.49	85.42
13	Eljanov, Pavel	2687	2732	2706.2	2718.9	16.25	2823	0.51	88.44	16.43	2854.3	0.22	38.3
14	Gelfand, Boris	2751	2753	2760.9	2751	16.92	2938.8	0.28	48.77	16.74	2908	0.37	64.36
15	Giri, Anish	2720	2746	2711.7	2710.1	16.41	2850	0.07	12.44	16.39	2847.1	0.02	3.47
16	Grischuk, Alexander	2764	2792	2780.6	2786	17.17	2982.3	0.38	65.32	17.23	2992.4	0.37	64.47
17	Ivanchuk, Vassily	2766	2753	2773.4	2771	17.12	2973.7	0.34	59.27	17.08	2966.6	0.36	62.56
18	Jakovenko, Dmitry	2741	2730	2734.6	2739.4	16.62	2887.5	0.42	73.42	16.69	2899.9	0.35	60.56
19	Kamsky, Gata	2762	2713	2776.1	2768.3	17.24	2994.5	0.46	80.05	17.15	2979.6	0.35	60.09
20	Karjakin, Sergey	2775	2770	2793.9	2796	17.39	3021.4	0.27	47.43	17.42	3026.5	0.22	38.82
21	Korobov, Anton	2702	2696	2701.1	2695.3	16.4	2848.9	0.57	99.42	16.39	2847.4	0.58	101.1
22	Kramnik, Vladimir	2795	2783	2809.5	2815.3	17.65	3066.3	0.59	102.97	17.73	3079.7	0.44	76.05
23	Leko, Peter	2732	2737	2733.4	2732.6	16.61	2885.8	0.55	96	16.77	2872.8	0.67	116.87
24	Le, QuangLiem	2705	2712	2725.2	2741.2	16.54	2873.6	0.17	28.76	16.54	2913.8	0.43	75.19
25	Malakhov, Vladimir	2709	2694	2704.5	2717	16.22	2817.8	0.38	66.59	16.42	2852.6	0.55	95.95
26	Mamedyarov, Shakhriyar	2764	2760	2765.8	2782.9	17	2953.2	0.84	145.7	17.22	2991.2	0.84	145.68
27	Moiseenko, Alexander	2710	2707	2683.9	2667.8	16.03	2784.8	0.57	98.66	15.84	2752.2	0.6	104.09
28	Morozevich, Alexander	2748	2719	2756	2718.9	16.95	2944.3	0.4	69.93	16.43	2853.7	0.4	68.97
29	Naiditsch, Arkadij	2708	2700	2686.3	2685.3	16.06	2790	0.46	80.26	16.03	2785.5	0.48	83.56
30	Nakamura, Hikaru	2760	2772	2789	2797.7	17.38	3019.2	0.35	60.45	17.5	3040.4	0.32	55.76
31	Navara, David	2710	2708	2702.2	2692.8	16.21	2816.2	0.33	57.49	16.09	2795.4	0.29	50.18
32	Nepomniachtchi, Ian	2707	2735	2720.3	2726.9	16.46	2858.9	0.17	29.39	16.58	2880.5	0.17	30.25
33	Ponomariov, Ruslan	2732	2723	2739.7	2743.5	16.62	2887.6	-0.06	-10.57	16.68	2897.8	-0.05	-8.97
34	Radjabov, Teimour	2793	2713	2785.4	2785.6	17.23	2993.2	0.12	20.87	17.2	2987.7	0.14	24.78
35	Riazantsev, Alexander	2705	2692	2699.1	2697.7	16.29	2829.4	0.15	25.31	16.29	2829.9	0.18	31.41
36	Shirov, Alexei	2708	2703	2729.4	2748.7	16.63	2889.6	0.61	105.45	16.88	2931.9	0.64	110.65
37	So, Wesley	2682	2731	2671.8	2674.6	15.89	2759.6	0.34	59.79	15.91	2763	0.18	32.05
38	Svidler, Peter	2747	2753	2762.3	2753.4	17	2953.3	0.18	31.5	16.87	2930	0.22	38.38
39	Tomashevsky, Evgeny	2725	2695	2725.5	2722.5	16.52	2869.2	0.56	96.61	16.48	2862.4	0.66	113.99
40	Topalov, Veselin	2771	2772	2730.5	2750.4	16.52	2869.9	0.58	101.18	16.76	2911.9	0.93	162.08
41	Vachier-Lagrave, Maxime	2711	2758	2712.1	2713.3	16.36	2842.8	0.37	63.85	16.38	2846.1	0.45	77.59
42	Vallejo Pons, Francisco	2694	2700	2686.8	2686.6	16.04	2786.3	0.49	85.02	16.02	2783.2	0.4	69.5
43	Vitiugov, Nikita	2694	2742	2708.5	2722.7	16.33	2837.6	0.47	81.58	16.54	2872.9	0.37	64.48
44	Wang, Hao	2737	2734	2746.4	2747	16.83	2924.5	0.39	67.78	16.87	2929.8	0.42	72.5
45	Wang, Yue	2696	2713	2707.9	2711.4	16.28	2827.3	0.42	72.76	16.31	2832.6	0.27	46.82
46	Wojtaszek, Radoslaw	2734	2724	2732.3	2731.2	16.64	2890.5	0.24	40.89	16.66	2894.8	0.26	45.06