Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 511–518
DOI: 10.15439/2014F33
ACSIS, Vol. 2

A Comparison between Different Chess Rating Systems for Ranking Evolutionary Algorithms

Niki Večeka*, Matej Črepinšeka*, Marjan Mernika* and Dejan Hrnčič†
*Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia
Email: {niki.vecek, marjan.mernik, matej.crepinsek}@uni-mb.si
†Adacta d.o.o., Maribor, Slovenia
Email: dejan.hrncic@adacta.si

***Abstract*—Chess Rating System for Evolutionary algorithms (CRS4EAs) is a novel method for comparing evolutionary algorithms which evaluates and ranks algorithms regarding the formula from the Glicko-2 chess rating system. It was empirically shown that CRS4EAs can be compared to the standard method for comparing algorithms - null hypothesis significance testing. The following paper examines the applications of chess rating systems beyond Glicko-2. The results of 15 evolutionary algorithms on 20 minimisation problems obtained using the Glicko-2 system were empirically compared to the Elo rating system, Chessmetrics rating system, and German Evaluation Number (DWZ). The results of the experiment showed that Glicko-2 is the most appropriate choice for evaluating and ranking evolutionary algorithms. Whilst other three systems' benefits were mainly the simple formulae, the ratings in Glicko-2 are proven to be more reliable, the detected significant differences are supported by confidence intervals, the inflation or deflation of ratings is easily detected, and the weight of individual results is set dynamically.**

***Index Terms*—chess rating system, ranking, evolutionary algorithms comparison, Glicko-2, Elo, Chessmetrics**

I. Introduction

A METHOD for comparing the algorithms is needed for determing whether one algorithm performs better than the other. As numerous effective evolutionary algorithms are appearing, a comparison with only one algorithm is now insufficient. This fact leads to the need for determing which of the multiple algorithms is better than the other. Which of them is the best and which the worst? A well-established method for comparing the experimental results of multiple evolutionary algorithms is Null Hypothesis Significance Testing (NHST) [22]. Whilst there are many variants of NHST, there are still some pitfalls regarding statistics and its application [2], [7], [13], [21] that imply that this field still needs attention. A novel method the Chess Rating System for Evolutionary Algorithms (CRS4EAs) [30] suggests using a chess rating system for evaluating the results and ranking the algorithms. CRS4EAs treats (i) evolutionary algorithms as chess players, (ii) one comparison between two algorithms as one game, and (iii) execution and evaluation of pairwise comparisons between all algorithms participating in the experiment as a tournament. Just like the standard comparison of two algorithms, one game in CRS4EA can have three outcomes: the first algorithm is better (and therefore wins), the second algorithm is better (and therefore wins), or they perform equally regarding predefined accuracy ϵ (they play a draw). It has been empirically shown that CRS4EAs is comparable to NHST, and can be used as a comparative method for evolutionary algorithms [30]. A CRS4EAs method is used within an open-source framework Evolutionary Algorithms Rating System (EARS) [8], [9]. CRS4EAs and EARS were developed to provide fairer and easier to understand comparisons between evolutionary algorithms. All the experiments in EARS are executed for the same number of optimisation problems, the algorithms are written in the same programming language (Java), have the same termination criteria, are initialised with the same random seed, and executed under the same hardware configuration. Hence, some factors that could affect the final results of the experiment were excluded [30]. The CRS4EAs uses the Glicko-2 chess rating system [18], since it is one of the newest and it consists of many preferences that look promising. In the proposed paper the Glicko-2 rating system is compared to three other better-known and well-established rating systems: Elo, Chessmetrics, and German Evaluation Number (DWZ). In order to compare these four rating systems the experiment was conducted for 15 evolutionary algorithms covering 20 minimisation problems. The analysis showed that comparing evolutionary algorithms the Glicko-2 was the most appropriate choice. One downside to the Glicko-2 is its complicated formulae, for the understanding of which mathematical and statistical knowledge is needed. The differences amongst players are more straightforward in the other three systems, however they are unsupported by any reliable measurements - they are arbitrary. Otherwise, Glicko-2 was shown to be more reliable: the detected significant differences are supported by a confidence interval, straightforward measurement for rating reliability, the control of conservativity/liberty is more correct, the weightings of individual results are set dynamically, improvement through time is considered in final results, the inflation or deflation of ratings is easily detected, and the selective pairing is not an issue. This paper presents the reasons why the first choice for rating system used in CRS4EAs was the Glicko-2.

The paper is structured as follows. Section II summarises four more popular chess rating systems. The formulae used in these systems are adapted for EARS and are used during the experiment. The CRS4EAs method and the experiment are introduced in Section III, followed by a detailed analysis of the obtained results. Section IV concludes the paper.

978-83-60810-58-3/$25.00

II. BACKGROUND

Chess is a strategic game of two players with three possible outcomes: the first player can win, the first player can lose, or the players can play a draw. Usually, the absolute power of a chess player is described using a number that is called a 'rating'. A player's rating is updated after each tournament the player participates in and each chess organisation has its own rating system with formulae that evaluate its players. In this section the more common chess rating systems are introduced. All the players are represented on the leaderboard, from best to worst and although there are different formulae behind updating the players' ratings, all of them have two things in common: a player's rating is always a positive integer and the player with the highest rating value is expected to be better. A player joins the tournament with k opponents in which the ith player has a rating R_i, and plays m games.

A. Elo

The best-known chess rating system is the Elo rating system [10] where the expected score of the ith player against the jth player is calculated using the formula in Eq. 1.

$$E(R_i, R_j) = \frac{1}{1 + 10^{(R_j - R_i)/400}} \tag{1}$$

The expected score of the ith against the jth player is the probability of i defeating j. Hence, the sum of the expected scores of the ith and jth players (against each other) equals 1. The score the ith player gained against the jth player is denoted by $S_{i,j}$ and equals 1 if the ith player won, 0 if ith player lost, or 0.5 for a draw. All the ratings are updated at the end of a tournament using the formula from Eq. 2. The new rating of the ith player is denoted by R'_i.

$$R'_i = R_i + K \sum_{j=1}^{m} \big(S_{i,j} - E(R_i, R_j)\big) \tag{2}$$

The K-factor is a constant that affect the emphasis of the difference between the actual score and the expected score. The USCF (United States Chess Federation) rating system implements the K-factor by dividing 800 by the sum of effective number of games a player's rating is based on (N_e) and the number of games the player completed during a tournament (m) [17]. Even though, the Elo system is famous for its simplicity and wide-usage, it has a few drawbacks such as properly setting the K-factor, an inaccurate distribution model, or unreliable rating.

B. Chessmetrics

The chess statistician Jeff Sonas proposed the usage of a more dynamic K-factor in his own chess rating system called Chessmetrics [27], described as 'a weighted and padded simultaneous performance rating'. Chessmetrics uses the following formula (Eq. 3) for updating the rating of the ith player.

$$R'_i = 43 + \frac{R_{per} * m + 4 * \sum_{j=1}^{k} R_j/k + 2300 * 3}{m + 7} \tag{3}$$

R_{per} is the performance rating calculated as $\sum_{j=1}^{k} R_j/k + (\sum_{j=1}^{m} S_{i,j}/m - 0.5) * 850$ and with the meaning that each 10% increase in percentage score corresponds to an 85 point advantage in the ratings [27].

C. German Evaluation Number (DWZ)

The simplest and one of the first rating systems was the Ingo rating system [20] by Anton Hoesslinger (1948), which has influenced many other rating system, including the Deutsche Wertungszahl (DWZ) [6]. DWZ is similar to the Elo rating system of the FIDE (World Chess Federation) but has improved in its own way since 1990 when it was first introduced. The expected score in DWZ is calculated using the same formula as the expected score in the Elo system (Eq. 1), whilst the rating is updated using the formula in Eq. 4.

$$R'_i = R_i + \sum_{j=1}^{m} \frac{800}{D + m} \big(S_{i,j} - E(R_i, R_j)\big) \tag{4}$$

D is the development coefficient (Eq. 5), dependent on the fundamental value D_0 (Eq. 6), the acceleration factor a (Eq. 7), and the breaking value b (Eq. 8).

$$\begin{aligned} D &= a * D_0 + b \\ 5 \leq D &\leq \begin{cases} \min(30, 5i) & \text{if } b = 0 \\ 150 & \text{if } b > 0 \end{cases} \end{aligned} \tag{5}$$

$$D_0 = \big(\frac{R_i}{1000}\big)^4 + J \tag{6}$$

The coefficient J differs according to the different ages of the players - the older the player, the bigger the J. The acceleration factor a (Eq. 7) cannot be higher than 1 or lower than 0.5, and is calculated only if a player younger than 20 years achieved more points than expected, otherwise a equals 1. The breaking value b (Eq. 8) is computed only if the player with a rating under 1300 achieved less points than expected, otherwise b equals 0.

$$a = \frac{R_i}{2000} \tag{7}$$

$$b = e^{\frac{1300 - R_i}{150}} - 1 \tag{8}$$

D. Glicko-2

One of main concerns about the Elo system is the possibility of a player winning the game and losing rating points, or losing the game and gaining rating points. Problems with unreliable ratings show in those games between players with the same rating, when one of them has not played for years and the other plays constantly - they would lose and gain the same amount of points. A less reliable rating is expected for the player who has not played in years, and a more reliable rating for the player who plays constantly. It is expected that if the first player wins his rating would go up more than the rating of the second player goes down. Because anything cannot be said about the player's gaming behaviour or the reliability of his power, Glickman [14] introduced a new chess rating system. The Glicko system [15] introduces a new value that represents

the reliability of a player's power - rating deviation RD - which is similar to standard deviation regarding statistics. RD_i is set to 350 at the beginning of the first tournament and updated (just as rating) at the end of each tournament. It decreases with each tournament the ith player participates in and increases with each tournament ith player skips. The maximum value of RD is 350, whilst the minimum is set by an organisation implementing the system (Glickman suggests 30). Rating deviation tells how reliable the player's rating is - the lower the RD the more reliable the rating. In 2012 Glickman updated its system and presented the Glicko-2 rating system [18], which is based on Glicko but has another variable that presents the reliability of the player's strength - rating volatility σ_i. The volatility indicates the degree of expected fluctuation in a player's rating. If σ_i is low the player performs at a consistent level, whilst high σ_i indicates erratic performances. Firstly, the rating R and rating deviation RD have to be converted from Glicko to the Glicko-2 rating system (Eq. 9).

$$\mu = \frac{R - 1500}{173.7178} \text{ and } \phi = \frac{RD}{173.7178} \tag{9}$$

The estimated variance v of the player's rating based only on game outcomes is calculated using the formula in Eq. 10.

$$v = \left(\sum_{j=1}^{m} g(\phi_j)^2 E(\mu_i, \mu_j, \phi_i)(1 - E(\mu_i, \mu_j, \phi_i)) \right)^{-1} \tag{10}$$

The gravity factor g (Eq. 11) and the expected score E (Eq. 12) are calculated using the following formulae.

$$g(\phi) = \frac{1}{\sqrt{1 + 3\phi^2/\Pi^2}} \tag{11}$$

$$E(\mu, \mu_i, \phi_i) = \frac{1}{1 + 10^{-g(\phi_i)(\mu - \mu_i)}} \tag{12}$$

Next, the estimated improvement in rating Δ (Eq. 13) has to be calculated where the pre-period rating μ_i is compared to the performance rating μ_j based only on the game outcomes $S_{i,j}$.

$$\Delta = v \sum_{j=1}^{m} g(\phi_j)(S_{i,j} - E(\mu_i, \mu_j, \phi_i)) \tag{13}$$

A new rating volatility σ' is found when using the Illinois algorithm [5] for a function $f(x) = \frac{e^x(\Delta^2 - \phi^2 - v - e^x)}{2(\phi^2 + v + e^x)^2} - \frac{x - \ln(\sigma^2)}{\tau^2}$ with accuracy of up to 6 decimal places. This method is used for finding zeros and once the zero x_0 of this function is found, σ' is set to $e^{x_0/2}$ and the pre-rating period value ϕ^* (Eq. 14) is calculated.

$$\phi^* = \sqrt{\phi^2 + \sigma'^2} \tag{14}$$

New values for rating deviation ϕ' (Eq. 15) and rating μ' (Eq. 16) are set.

$$\phi' = \frac{1}{\sqrt{\frac{1}{(\phi^*)^2} + \frac{1}{v}}} \tag{15}$$

$$\mu' = \mu + \phi' \sum_{i=1}^{m} g(\phi_i)(S_i - E(\mu, \mu_i, \phi_i)) \tag{16}$$

Finally, the new rating R' and new rating deviation RD' are converted from the Glicko-2 to the Glicko system using the formulae in Eq. 17.

$$R' = 173.7178\mu' + 1500 \text{ and } RD' = 173.7178\phi' \tag{17}$$

All of these systems have their own advantages (Table I), however, Glicko-2 contains most of them despite its more complicated formula (in comparison with other systems).

TABLE I: Preferences a chess rating contains.

Preference	Elo	Chessmetrics	DWZ	Glicko-2
Simple formula	✔	✔	✔	
Player's age influence			✔	
Dynamic weight factor		✔	✔	✔
Control over selective pairing				✔
Time varying impact				✔
Bayesian approach				✔
Straightforward measurement of rating inflation and deflation				✔
Straightforward measurement of rating reliability				✔
Straightforward measurement of differences between ratings				✔

Our implementations of these four algorithms were used in the following experiment.

III. Experiment

This experiment was conducted using the novel method for comparing and ranking the evolutionary algorithms CRS4EAs [30]. The experiment in CRS4EAs is executed as any other experiment, however each outcome of each algorithm regarding every optimisation problem must be saved for further comparison. In the CRS4EAs the run-by-run comparison the roles of the chess players adopt evolutionary algorithms. Each outcome in every run for every optimisation problem of one algorithm is compared to the corresponding outcome of the other algorithm. Such a comparison is called one 'game'. If the difference between compared outcomes is less than the predefined ϵ, the final score of this game is a draw, otherwise the algorithm with the outcome closer to the optimum of the optimisation problem wins. With k algorithms ($k - 1$ opponents), N optimisation problems, and n runs, one algorithm plays $n * N * (k - 1)$ games during one tournament. Hence, in our tournament $n * N * k * (k - 1)/2$ games are played. The whole process is presented in the flowchart in Fig. 1. The chess rating system used in CRS4EAs is Glicko-2, however due to this being an experiment, other chess rating systems were implemented as well.

In the presented experiment our implementations of $k = 15$ evolutionary algorithms were compared for $N = 20$ optimisation problems over $n = 100$ runs. The simplest algorithm used in the experiment was the basic random search (RWSi) [24]. Next being Teaching Learning Based Optimization (TLBO) [3], [25]. There were two variants of evolutionary strategies (ES(1+1) and CMA-ES) [19], [26], 10 variants of the Differential Evolution [4], [23], [29], [31], and the Self-adaptive

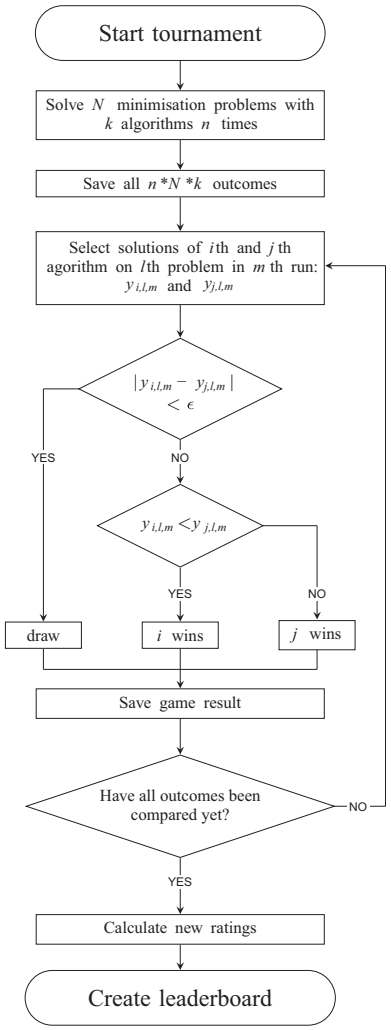


Fig. 1: Flowchart of experiment's execution in CRS4EAs.

Differential Evolution (jDE) [1]. The optimisation problems were from the Special Session and Competition on Large-Scale Global Optimization CEC 2010 [28]. The termination critera for each algorithm was maximum number of evaluations $Max\_FEs = 10^5$. The threshold for a draw was set at $\epsilon = 1.0e-06$, and the initial rating was set to 1500 for each rating system to provide a fairer comparison. Detailed descriptions of the algorithms and optimisation problems can be found in [30]. Other properties for the rating systems can be seen in Table II. R_{init} represents the initial rating for a new player, rating intervals and rating ranges present the values for detecting the differences in the powers of the algorithms, K is the K-factor in Elo, N_e is the number of games a player's rating is based on, m is the number of games the player completed during a tournament, J is the age coefficient from Chessmetrics, RD_{init} is the initial rating deviation for a new player, RD_{min} is the minimum rating deviation, and RD_{max} is the maximum rating deviation. Whilst Glicko-2 uses the straightforward values for detecting significant differences - R, RD, and rating interval - other systems do not consist of such preferences. Two algorithms are significantly different if their rating intervals do not overlap. In Glicko-2 the 99.7% rating (confidence) interval is defined by $[R-3RD, R+3RD]$. The rating range that distinguishes between the powers of two players in Elo equals 200 rating points. The minimum Elo rating can be 100 points, then the players are classified in categories: 100-199 points (J), 200-399 points (I), 400-599 points (H), 600-799 points (G), 800-999 points (F), 1000-1199 points (E), 1200-1399 points (D), 1400-1599 points (C), 1600-1799 points (B), 1800-1999 points (A), 2000-2199 points (expert), etc. The same is done for DWZ and Chessmetrics, but while DWZ uses the same categories as Elo, the Chessmetrics' categories differ by 100 points. However, it must be explicitly pointed out that this classification of categories is not a straighforward way of detecting significant differences as the confidence intervals in Glicko-2.

TABLE II: Properties for the chess ratings used during the experiment.

Chess rating system	Properties
Elo	$R_{init} = 1500$
	Rating range 200 points
	$K = \frac{800}{N_e+m}$
Chessmetrics	$R_{init} = 1500$
	Rating range 100 points
DWZ	$R_{init} = 1500$
	$J = 15$ for all players
	Rating range 200 points
Glicko-2	$R_{init} = 1500$
	$RD_{init} = 350$, $RD_{min} = 50$, $RD_{max} = 350$
	Rating interval $[R-3RD, R+3RD]$

The ratings evolutionary algorithms obtained for each rating system are presented on a group leaderboard in Table III. All the algorithms obtained minimum rating deviations of 50 points in the Glicko-2 system. Although, different formulae were used in different chess rating systems, the orders of the ratings were almost the same. The only rating system for which the order of algorithms was different was Elo where CMA-ES, DE/best/2/exp and DE/rand/1/exp go in reverse order. These three algorithms, however, are really close regarding rating points. In order to obtain a better picture the average ranking of the algorithms by data sets, i.e. Friedman ranking [11], [12] was added in the last column. A statistical analysis and comparison with NHST can be found in [30]. All the obtained ratings are displayed in Fig. 2 where distributions of ratings for each rating system are shown. Maximum and minimum overall rating values were obtained in Elo. These ratings were more scattered and there was a big gap (435 points) between the 7th algorithm DE/best/1/exp and the 8th algorithm DE/best/2/bin by dividing the algorithms into two groups: algorithms from 1 to 7 and algorithms from 7 to 15. The Chessmetrics and DWZ ratings seemed to be equally distributed, but the difference between the corresponding rating points varied between 20 to 59 points. The difference was bigger for better performing algorithms

TABLE III: Leaderboard with ratings the algorithms obtained using four different rating systems and the average ranking (AR) of the algorithms by data sets, i.e. Friedman ranking [11], [12].

i	Algorithm	Elo	Chessmetrics	DWZ	Glicko-2	AR
1	JDE/rand/1/bin	2014	1812	1753	1829	3.6
2	DE/rand/2/exp	1996	1772	1715	1779	3.425
3	CMA-ES	**1972**	1767	1711	1774	4.325
4	DE/best/2/exp	**1982**	1761	1705	1766	4.675
5	DE/rand/1/exp	**1985**	1758	1702	1762	4.325
6	DE/rand/2/bin	1940	1704	1651	1696	4.75
7	DE/best/1/exp	1890	1626	1578	1602	7.675
8	DE/best/2/bin	1455	1588	1542	1554	7.975
9	DE/rand-to-best/1/exp	1361	1575	1530	1540	7.05
10	DE/rand/1/bin	1221	1516	1475	1467	8.5
11	DE/rand-to-best/1/bin	1129	1375	1342	1294	10.8
12	TLBO	1078	1297	1268	1199	12.05
13	DE/best/1/bin	1057	1268	1241	1164	12.55
14	RWSi	1000	1178	1156	1054	13.7
15	ES(1+1)	983	1151	1131	1020	14.6

(59 for JDE/rand/1/bin) and smaller for worse performing algorithms (20 for ES(1+1)). The biggest gap in ratings for Glicko-2, DWZ, and Chessmetrics was between the 10th algorithm DE/rand/1/bin and the 11th algorithm DE/rand-to-best/1/bin. Algorithms DE/rand/2/exp ($i = 2$), CMA-ES ($i = 3$), DE/best/2/exp ($i = 4$), and DE/rand/1/exp ($i = 5$) were close in ratings for all four rating systems.

An interesting outlook regarding the results of a tournament is when examinating wins, losses, and draws (Table IV). This is not only useful in chess but also in comparison with evolutionary algorithms. The number of wins, losses, and draws can tell a lot about how one algorithm performed against another. For example, JDE/rand/1/bin was the overall best algorithm - it had the most wins and the least losses - but when comparing its performance with the performance of the worst algorithm ES(1+1) - with the least wins and the most losses - showed that ES(1+1) defeated JDE/rand/1/bin in 1 out of 2000 (=20*100) games. It could be concluded that the JDE/rand/1/bin performed with outliers as this is a phenomenon that is also detected with other worse algorithms: DE/rand-to-best/1/bin (2 outliers), TLBO (2 outliers), DE/best/1/bin (2 outliers), and RWSi (2 outliers). An interesting fact is that CMA-ES has more wins than DE/rand/2/exp but is ranked one place lower. This is due to the fact that CMA-ES also has more loses and less draws. However, as mentioned before the difference in ratings is small. Table IV also shows that the draws were less common in those games with low-ranked algorithms - even between the low-ranked algorithms themselves. The draws were fairly common in games between the first half of the algorithms, whilst in games with algorithms that were ranked lower than 8th place the draws hardly appeared. The most draws (1112) were played between DE/rand/2/exp and DE/rand/1/exp. DE/rand/2/exp, DE/rand/2/bin, and DE/rand-to-best/1/exp were the only three algorithms that won the absolute number of games (2000) against at least one opponent. DE/rand/2/exp won absolutely against TLBO, DE/best/1/bin, RWSi, and ES(1+1), DE/rand/2/bin against RWSi, and DE/rand-to-best/1/exp against ES(1+1).

The detected significant differences are shown in Fig. 3. As Chessmetrics has the lowest threshold for classifying players into groups (100 rating points), the highest distinctions (90) between players were detected within this system. Elo and DWZ had the same threshold (200 rating points), but more distinctions were detected in Elo, due to the fact that the obtained players' ratings in Elo had wider ranges. Chessmetrics detected 10 differences more than DWZ, 8 differences more than Elo, and there was no difference in those detected by DWZ or Elo and those Chessmtrics was not. DWZ detected 8 differences that Elo did not, and Elo detected 11 differences that DWZ did not. These differences are listed in Table V.

TABLE V: System marked with ✓ detected differences in the ratings of the listed algorithms, whilst the system marked with ✗ did not.

Chessmetrics ✓ DWZ ✗	Chessmetrics ✓ Elo ✗
JDE/rand/1/bin vs. DE/rand/2/exp	DE/rand/2/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. CMA-ES	CMA-ES vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/best/2/exp	DE/best/2/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/rand/1/exp	DE/rand/1/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/rand/2/bin	DE/rand/2/bin vs. DE/best/1/exp
DE/best/1/exp vs. DE/best/2/bin	DE/rand-to-best/1/bin vs. TLBO
DE/best/1/exp vs. DE/rand-to-best/1/exp	DE/rand-to-best/1/bin vs. DE/best/1/bin
DE/best/1/exp vs. DE/rand/1/bin	DE/rand-to-best/1/bin vs. RWSi
DE/rand-to-best/1/bin vs. TLBO	
DE/rand-to-best/1/bin vs. DE/best/1/bin	

DWZ ✓ Elo ✗	Elo ✓ DWZ ✗
DE/rand/2/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/2/exp
CMA-ES vs. DE/best/1/exp	JDE/rand/1/bin vs. CMA-ES
DE/best/2/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/best/2/exp
DE/rand/1/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/1/exp
DE/rand/2/bin vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/2/bin
DE/rand-to-best/1/bin vs. RWSi	DE/best/1/exp vs. DE/best/2/bin
TLBO vs. RWSi	DE/best/1/exp vs. DE/rand-to-best/1/exp
DE/best/1/bin vs. RWSi	DE/best/1/exp vs. DE/rand/1/bin
	DE/best/2/bin vs. DE/rand-to-best/1/exp
	DE/best/2/bin vs. DE/rand/1/bin
	RWSi vs. ES(1+1)

It appears that Elo, Chessmetrics, and DWZ are more liberal, and the conservativity could be increased with a wider rating range between categories. However controlling the conservativity in such way would not be as efficient as in Glicko-2 where conservativity is controlled by setting the minimal rating deviation and choosing an appropriate confidence interval. In Glicko-2 the algorithms' ratings were compared pairwisely, whilst with the other three systems algorithms were classified into groups and then compared regarding them. Also, the significances of the differences detected within Elo, Chessmetrics, and DWZ are unknown, as there was no statistical tool for measuring them and the choice of rating range is arbitrary. On the other hand, Glicko-2 detected 50 significant differences that were made with 99.7% confidence and were comparable to NHST [30]. The tests of significance used for NHST analysis were the Friedman

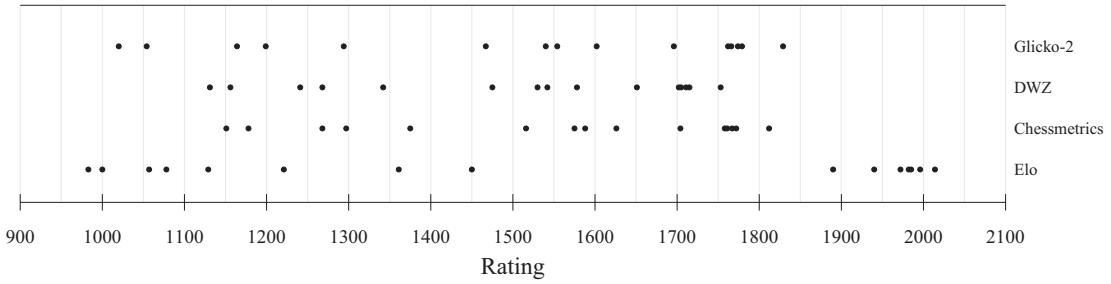


Fig. 2: Distributions of the ratings for all 4 rating systems.

TABLE IV: Wins (W), losses (L), and draws (D) algorithms obtained in tournaments against each other. When the number of losses is relatively small (e.g., < 75) the occurance of outliers is very likely.

i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σ
	W	/	642	769	750	775	747	1417	1389	1919	1976	1998	1998	1998	1998	1999	20375
1 JDE/rand/1/bin	L	/	423	659	407	365	397	64	245	67	24	2	2	2	2	1	2660
	D	/	935	572	843	860	856	519	366	14	0	0	0	0	0	0	4956
	W	423	/	711	519	474	624	1182	1299	1690	1843	1993	2000	2000	2000	2000	18758
2 DE/rand/2/exp	L	642	/	665	502	414	380	298	334	305	157	7	0	0	0	0	3704
	D	935	/	624	979	1112	996	520	367	5	0	0	0	0	0	0	5538
	W	659	665	/	765	688	853	1295	1371	1662	1843	1901	1910	1942	1898	1968	19420
3 CMA-ES	L	769	711	/	593	700	559	246	164	334	157	99	90	58	102	32	4614
	D	572	624	/	642	612	606	459	465	4	0	0	0	0	0	0	3984
	W	407	502	593	/	570	684	1175	1314	1633	1774	1992	1982	1999	1998	1999	18622
4 DE/best/2/exp	L	750	7519	765	/	547	462	309	306	356	226	8	18	1	2	1	11270
	D	843	979	642	/	883	854	516	380	11	0	0	0	0	0	0	5108
	W	365	414	700	547	/	555	1157	1244	1713	1846	1969	1976	1987	1984	1996	18453
5 DE/rand/1/exp	L	775	474	688	570	/	537	353	390	279	154	31	24	13	16	4	4308
	D	860	1112	612	883	/	908	490	366	8	0	0	0	0	0	0	5239
	W	397	380	559	462	537	/	1028	1080	1525	1584	1671	1905	1759	2000	1851	16738
6 DE/rand/2/bin	L	747	624	853	684	555	/	460	552	471	416	329	95	241	0	149	6176
	D	856	996	606	854	908	/	512	368	4	0	0	0	0	0	0	5104
	W	64	298	246	309	353	460	/	1114	1103	1333	1891	1924	1987	1977	1993	15052
7 DE/best/1/exp	L	1417	1182	1295	1175	1157	1028	/	530	893	667	109	76	13	23	7	9572
	D	519	520	459	516	490	512	/	356	4	0	0	0	0	0	0	3376
	W	245	334	164	306	390	552	530	/	979	1097	1691	1862	1985	1997	1996	14128
8 DE/best/2/bin	L	1389	1299	1371	1314	1244	1080	1114	/	1017	903	309	138	15	3	4	11200
	D	366	367	465	380	366	368	365	/	4	0	0	0	0	0	0	2681
	W	67	305	334	356	279	471	893	1017	/	1352	1982	1987	1999	1999	2000	15041
9 DE/rand-to-best/1/exp	L	1919	1690	1662	1633	1713	1525	1103	979	/	648	18	13	1	1	0	12905
	D	14	5	4	11	8	4	4	4	/	0	0	0	0	0	0	54
	W	24	157	157	223	154	416	667	903	648	/	1893	1930	1970	1974	1990	13106
10 DE/rand/1/bin	L	1976	1843	1843	1774	1846	1584	1333	1097	1352	/	107	70	30	26	10	14891
	D	0	0	0	0	0	0	0	0	0	/	0	0	0	0	0	0
	W	2	7	99	8	31	329	109	309	18	107	/	1618	1846	1976	1993	8452
11 DE/rand-to-best/1/bin	L	1998	1993	1901	1992	1969	1671	1891	1691	1982	1893	/	381	153	24	7	19546
	D	0	0	0	0	0	0	0	0	0	0	/	1	1	0	0	2
	W	2	0	90	18	24	95	76	138	13	70	381	/	1199	1967	1809	5882
12 TLBO	L	1988	2000	1910	1982	1976	1905	1924	1862	1987	1930	1618	/	801	33	191	22107
	D	0	0	0	0	0	0	0	0	0	0	1	/	0	0	0	1
	W	2	0	58	1	13	241	13	15	1	30	153	801	/	1770	1834	4932
13 DE/best/1/bin	L	1998	2000	1942	1999	1987	1759	1987	1985	1999	1970	1846	1199	/	230	166	23067
	D	0	0	0	0	0	0	0	0	0	0	1	0	/	0	0	1
	W	2	0	102	2	16	0	23	3	1	26	24	33	230	/	1501	1963
14 RWSi	L	1998	2000	1898	1998	1984	2000	1977	1997	1999	1974	1976	1967	1770	/	499	26037
	D	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0	0
	W	1	0	32	1	4	149	7	4	0	10	7	191	166	499	/	1071
15 ES(1+1)	L	1999	2000	1968	1999	1996	1851	1993	1996	2000	1990	1993	1809	1834	1501	/	26929
	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0

and Nemenyi tests with critical difference $CD = 4.79$. The first implied that there are significant differences between algorithms, and the other found 43 significant differences that were similar to those found with Glicko-2 (Fig. 3e). The executed experiment therefore showed that the Glicko-2 rating system is more appropriate for comparison and ranking of evolutionary algorithms. It provides more reliable ratings and more evident way of detecting significant differences. Hence, the preferences of the Glicko-2 (Table I) do not only contribute in tournaments between chess players but also in comparison between evolutionary algorithms.

IV. Conclusion

This paper conducted a comparison of four chess rating systems for ranking evolutionary algorithms. All the rating systems were implemented within EARS software, executed as an experiment, and analysed using the CRS4EAs method. The experiment showed that the Glicko-2 rating system is the most appropriate for ranking evolutionary algorithms. The main reason lies in the detection of significant differences amongst players and the formation of a confidence interval that allows direct comparison with null hypothesis significance testing. The other three systems - Elo, Chessmetrics, and DWZ - use simpler methods for detecting differences between ratings. Players are classified into categories and the differences in powers depend on the category the player belongs to. A new method for detecting the differences between players could increase the efficiencies of these systems, if the proposed method were dynamic (similar to Glicko-2). Otherwise, the results obtained from small tournaments (with a small number of algorithms or a small number of optimisation problems) would be unreliable. The conservativity/liberty of the method can be more efficiently controlled within Glicko-2. Elo, Chessmetrics, or DWZ can be improved by using some factors that are important for chess players (e.g., a player's age or the colour of pieces), but are irrelevant when comparing evolutionary algorithms. The results in CRS4EAs can be examined by observing the number of wins, losses, and draws amongst different players. Using this approach the outliers can be detected and the number of draws can indicate which algorithms are more likely to play a draw. In this paper we have empirically shown that various chess rating systems can be used for comparison amongst evolutionary algorithms and their rankings. The rationale as to why Glicko-2 may be a better choice than other chess systems for comparing the evolutionary algorithms has also been discussed in details. In the future, we will continue using Glicko-2 for CRS4EAs, with more focus on tuning the parameters. Glicko-2 was proven to be more reliable and dynamic than other older systems.

References

[1] J. Brest, S. Greiner, B. Bošković, M. Mernik, V. Žumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation*, 10(6):646–657, 2006.

[2] J. Cohen. The earth is round ($p < .05$). *American psychologist*, 49(12):997–1003, 1994.

[3] M. Črepinšek, S.H. Liu, L. Mernik. A Note on Teaching-Learning-Based Optimization Algorithm. *Information Sciences*, 212:79–93, 2012.

[4] S. Das, P.N. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31, 2011.

[5] M. Dowell, P. Jarratt. A modified regula falsi method for computing the root of an equation. *BIT Numerical Mathematics*, 11(2):168–174, 1971.

[6] Deutscher Schachbund [Online]. Available: http://www.schachbund.de/wertungsordnung.html

[7] T. Dyba, V.B. Kampenes, D.I. Sjoberg. A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8):745–755, 2006.

[8] Evolutionary Algorithms Rating System [Online]. Available: http://earatingsystem.appspot.com/

[9] Evolutionary Algorithms Rating System (Github) [Online]. Available: https://github.com/matejxxx/EARS

[10] A.E. Elo. The rating of chessplayers, past and present (Vol. 3). *Batsford*, 1978.

[11] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

[12] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[13] J. Gill. The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3):647–674, 1999.

[14] M.E. Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3:59–102, 1995.

[15] M.E. Glickman. The glicko system. *Boston University*, 1995.

[16] M.E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999.

[17] M.E. Glickman. Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6):673–689, 2001.

[18] M.E. Glickman. Example of the Glicko-2 system. *Boston University*, 2012.

[19] N. Hansen. The CMA Evolution Strategy: A Comparing Review. *Towards a new evolutionary computation, Springer*, 75–102, 2006.

[20] D. Hooper, K. Whyld. The Oxford Companion to Chess. Oxford University Press, 1992.

[21] B.A. Kitchenham, S.L. Pfleeger, L.M. Pickard, P.W. Jones, D.C. Hoaglin, K. El Emam, J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, 2002.

[22] J. Neyman, E. Pearson. On the problem of the most efficient test of statistical hypothesis. *Philosophical Transaction of the Royal Society of London - Series A*, 231:289–337, 1933.

[23] M.G. Epitropakis, V.P. Plagianakos, M.N. Vrahatis. Balancing the exploration and exploitation capabilities of the differential evolution algorithm. *IEEE World Congress on Computational Intelligence 2008*, 2686–2693, 2008.

[24] L.A. Rastrigin. The convergence of the random search method in the extremal control of a many-parameter system. *Automation and Remote Control*, 24(10):1337–1342, 1963.

[25] R.V. Rao, V.J. Savsani, D.P. Vakharia. Teaching-Learning-Based Optimization: An optimization method for continuous non-linear large scale problems. *Information Sciences*, 183(1):1–15, 2012.

[26] I. Rechenberg. Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. *Frommann-Holzboog*, 1973.

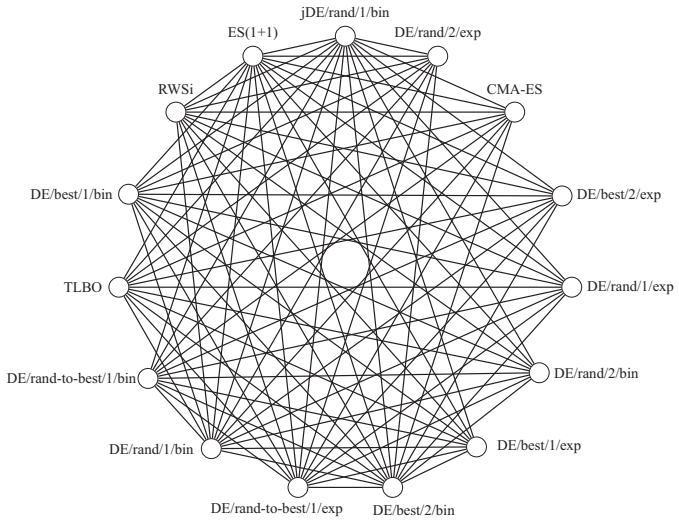
[27] J. Sonas. http://www.chessmetrics.com, Februar 2014.

[28] K. Tang, X. Li, P.N. Suganthan, Z. Yang, T. Weise. Benchmark Functions for the CEC2010 Special Session and Competition on Large-Scale Global Optimization. *Nature Inspired Computation and Applications Laboratory*, 2009.

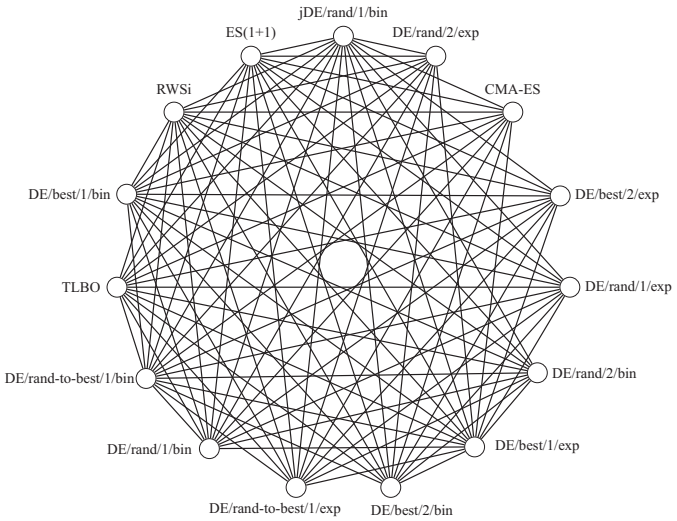
[29] J. Tvrdik. Adaptive differential evolution: application to nonlinear regression. *In Proceedings of the International Multiconference on Computer Science and Information Technology*, 193–202, 2007.

[30] N. Veček, M. Mernik, M. Črepinšek. A Chess Rating System for Evolutionary Algorithms - A New Method for the Comparison and Ranking of Evolutionary Algorithms. *Information Sciences*, 277:656-679, 2014.

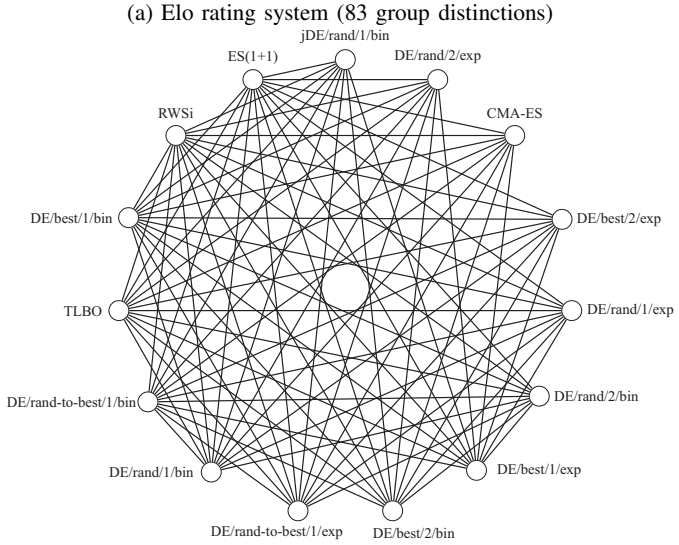
[31] D. Zaharie. A comparative analysis of crossover variants in differential evolution. *Proceedings of IMCSIT*, 171-181, 2007.



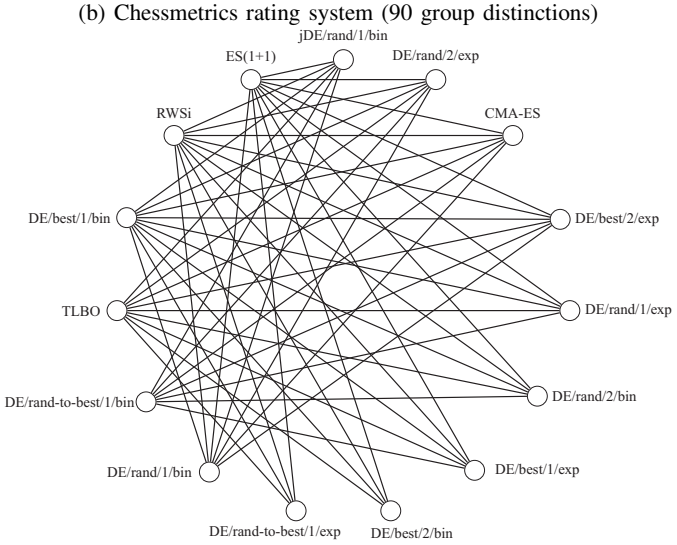
(a) Elo rating system (83 group distinctions)



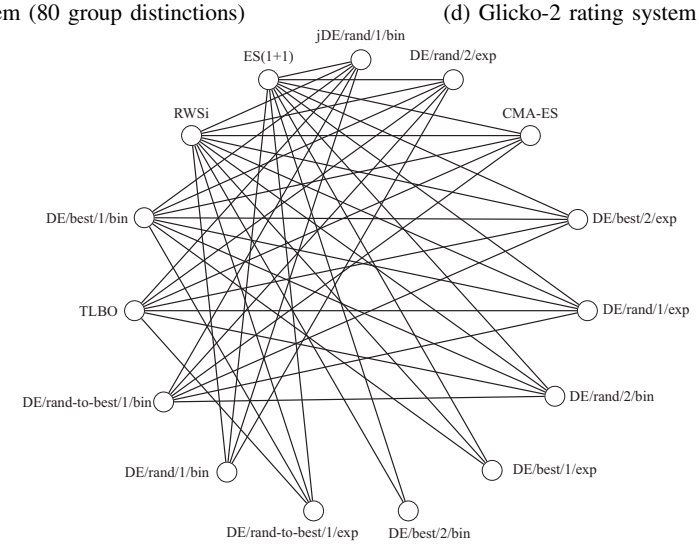
(b) Chessmetrics rating system (90 group distinctions)



(c) DWZ rating system (80 group distinctions)



(d) Glicko-2 rating system (50 significant differences)



(e) Null Hypothesis Significance Testing (43 significant differences)

Fig. 3: Detected differences amongst four rating systems. Two algorithms are connected when they are not within the same rating group (Fig. 3a, 3b, 3c) or are significantly different with probability 99.7% (Fig. 3d) or are significantly different with Null Hypothesis Significance Testing - Friedman test and $CD = 4.79$ (Fig. 3e).