Understanding Distributions of Chess Performances

Kenneth W. Regan[1], Bartlomiej Macieja[2], and Guy McC. Haworth[3]

[1] Department of CSE, University at Buffalo, Amherst, NY 14260 USA; regan@buffalo.edu
[2] Warsaw, Poland
[3] School of Systems Engineering, University of Reading, UK; guy.haworth@bnc.oxon.org

Abstract. This paper presents evidence for several features of the population of chess players, and the distribution of their performances measured in terms of Elo ratings and by computer analysis of moves. Evidence that ratings have remained stable since the inception of the Elo system in the 1970's is given in several forms: by showing that the population of strong players fits a simple logistic-curve model without inflation, by plotting players' average error against the FIDE category of tournaments over time, by skill parameters from a model that employs computer analysis keeping a nearly constant relation to Elo rating across that time, and by the same model showing steady improvement in its skill measures since the dawn of organized chess. The distribution of the model's *Intrinsic Performance Ratings* can hence be used to compare populations that have limited interaction, such as between players in a national chess federation and FIDE, and ascertain relative drift in their respective rating systems.

Keywords. Computer games, chess, ratings, statistics.

1 Introduction

Chess players form a dynamic population of varying skills, fortunes, and aging tendencies, and participate in zero-sum contests. A numerical rating system based only on the outcomes of the contests determines everyone's place in the pecking order. There is much vested interest in the accuracy and stability of the system, with significance extending to other games besides chess and potentially wider areas. Several fundamental questions about the system lack easy answers: How accurate are the ratings? How can we judge this? Have ratings inflated over time? How can different national rating systems be compared with the FIDE system? How much variation in performance is intrinsic to a given skill level? Answering the last question may inform the question of the optimal "K-factor" governing the rate of change of ratings.

This paper seeks statistical evidence beyond previous direct attempts to measure the system's features. We examine player rating distributions across time since the inception of the Elo rating system by the World Chess Federation (FIDE) in 1971, and extend before it, showing numbers that differ markedly from the "ChessMetrics" system of Sonas [Son11]. We continue work by Haworth, DiFatta, and Regan [Haw03,Haw07,DHR09,RH11] on measuring performances 'intrinsically' by the quality of moves chosen rather than the results of games. The models in this work have adjustable parameters that correspond to skill levels calibrated to the Elo scale. We have also measured aggregate error rates judged by computer analysis of entire tournaments, and plotted them against the Elo rating *category* of the tournament. Major findings of this paper extend the basic result of [RH11] that ratings have remained stable since the 1970's, contrary to the popular wisdom of extensive "rating inflation."

2 Ratings and Distributions

The Elo rating system, which originated for chess but is now used by many other games and sports, provides rules for updating ratings based on performance in games against other Elo-rated players, and for bringing new (initially 'unrated') players into the system. In chess they have a numerical scale where 2800 is achieved by a handful of top players today, 2700 is needed for most highest-level tournament invitations, 2600 is a 'strong' grandmaster (GM), while 2500 is typical of most GM's, 2400 of International Masters, 2300 of FIDE Masters, and 2200 of masters in national federations.

We emphasize that the ratings serve two primary purposes:

1. To indicate position in the world ranking.
2. To indicate a level of skill.

These two purposes lead to different interpretations of what it means for "inflation" to occur. According to view 1. 2700 historically meant what the neighborhood of 2800 means now: being among the very best, a true world championship challenger. As late as 1981, Anatoly Karpov topped the ratings at 2695, so no one had 2700, while today there are forty-five players 2700 and higher, some of whom have never been invited to an elite event. Under this view, inflation has occurred *ipso-facto*.

While view 2. is fundamental and has always had adherents, for a long time it had no reliable benchmarks. The rating system itself does not supply an intrinsic meaning for the numbers and does not care about their value: arbitrarily add 1000 to every figure in 1971 and subsequent initialization of new players, and relative order today would be identical. However, recent work [DHR09,**?**] provides a benchmark to calibrate the Elo scale to games analyzed in the years 2006–2009, and finds that ratings fifteen and thirty years earlier largely correspond to the same benchmark positions. In particular, today's echelon of over forty 2700+ players all give the same or better benchmarks than Karpov and Viktor Korchnoi in their prime. We consider that two further objections to view 2. might take the following forms:

(a) If Karpov and Korchnoi had access to today's computerized databases and more extensive opening publications, they would have played 50 to 100 points higher—as Kasparov did as the 1980's progressed.
(b) Karpov and Korchnoi were supreme strategists whose *depth* of play does not show up in ply-limited computer analysis.

We answer (a) by saying we are concerned only with the quality of moves made on the board, irrespective of their provenance. Regarding also (b) we find that today's elite make fewer clear mistakes than their forbears. This factor impacts skill apart from strategic depth. The model from [RH11] used in this paper by its nature weights the relative importance of avoiding mistakes.

Our position in subscribing to view 2. is summed up as *today's players deserve their ratings*. The numerical rating should have a fixed meaning apart from giving a player's rank in the world pecking order. In subsequent sections we present the following evidence that there has been no inflation, and that the models used for our conclusions produce reasonable distributions of chess performances.

- The proportion of Master-level ratings accords exactly with what is predicted from the growth in population alone.
- A version of the "average difference" (AD) statistic used by Guid and Bratko [GB06] to compare world championship matches shows that tournaments of a given category have seen fairly constant AD over time.
- "Intrinsic Ratings" as judged from computer analysis have likewise remained relatively constant as a function of Elo rating over time—for this we refine the method of Regan and Haworth [RH11].
- Intrinsic Ratings for the world's top players have increased steadily since the mid-1800s, mirroring the way records have improved in many other sports and human endeavors.
- Intrinsic Performance Ratings (IPR's) for players in events fall into similar distributions as assumed for Tournament Performance Ratings (TPR's) in the rating model. They can also judge inflation or deflation between two rating systems, such as those between FIDE and a national federation much of whose population has little experience in FIDE-rated events.

The last item bolsters the the Regan-Haworth model [RH11] as a reliable indicator of performance, and hence enhances the significance of the third and fourth items.

The persistence of rating controversies after many years of the standard analysis of rating curves and populations calls to mind the proverbial elephant that six blind men are trying to picture. When our five non-standard modes of analysis are agreeing, however, they can be said to have gained a reasonable understanding of the elephant after all. Besides providing new insight into distributional analysis of chess performances, the general nature of our tools allows application in other games and fields besides chess.

3 Population Statistics

Highlighted by the seminal work of de Solla Price on the metrics of science [dSP61], researchers have gained an understanding of the growth of human expertise in various subjects. In an environment with no limits on resources for growth, de Solla Price showed that the rate of growth is proportional to the population,

$$\frac{dN}{dt} \sim aN, \tag{1}$$

which yields an exponential growth curve. For example, this holds for a population of academic scientists, each expected to graduate some number $a > 1$ of students as new academic scientists. However, this growth cannot last forever, as it would lead to a day when the projected number of scientists would be greater than the total world population. Indeed, David Goodstein [Goo94] showed that the growth of PhD's in physics produced each year in the United States stopped being exponential around 1970, and now remains at a constant level of about 1000.

The theory of the growth of a population under limiting factors has been successful in other subjects, especially in biology. Since the work of Pierre-Franois Verhulst [Ver38] it has been widely verified that in an environment with limited resources the growth of animals (for instance tigers on an island) can be well described by a logistic function

$$N(t) = \frac{N_{max}}{(1 + a(exp)^{-bt})} \qquad \text{arising from} \qquad \frac{dN}{dt} \sim aN - bN^2, \tag{2}$$

where bN^2 represents a part responsible for a decrease of a growth due to an over-population, which is quadratic insofar as every animal interacts, for instance fights for resources, with every other animal. We demonstrate that this classic model also describes the growth of the total number of chess players in time with a high degree of fit.

We use a minimum rating of 2203—which FIDE for the first three Elo decades rounded up to 2205—because the rating floor and the start rating of new players have been significantly reduced from 2200 which was used for many years.

Figure 1: Growth of number of players rated at least 2203 since 1971

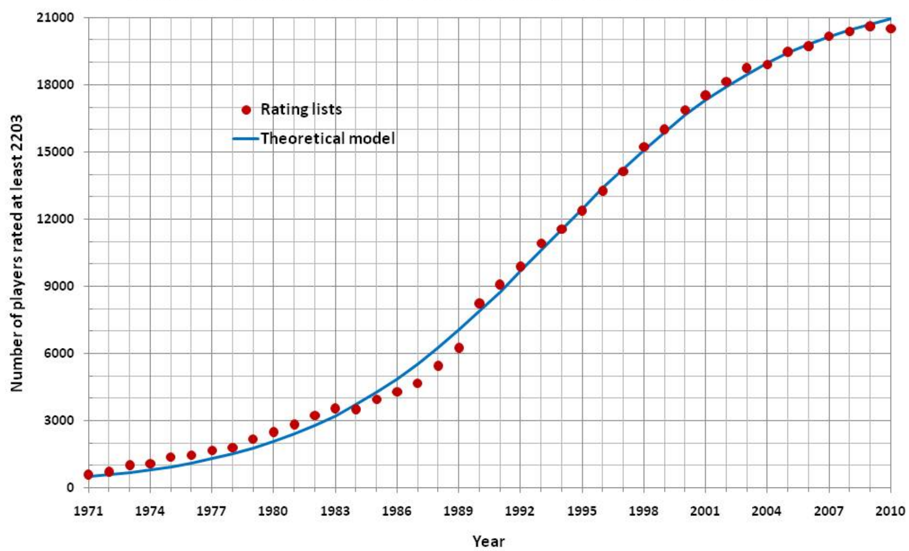


Figure 1 shows the number of 2203+ rated players, and a curve obtained for some particular values of a, b, and N_{max}. Since there are many data points and only three parameters, the fit is striking. This implies that the growth of the number of chess players can be explained without a need to postulate inflation.

The model makes a testable prediction. The best fit is obtained for N_{max} approximately 22,250. Having in mind that the total number of players rated at least 2203 is already almost 21,000, we do not expect a further significant growth in the number of master-level players worldwide.

4 Average Error and Results by Tournament Categories

The first author has run automated analysis of almost every major event in chess history, using the program RYBKA 3 [RK07] to fixed reported depth 13 ply[4] in Single-PV mode. This mode is similar to how Matej Guid and Ivan Bratko [GB06] operated the program CRAFTY to depth (only) 12, and how others have run other programs since. Game turns 1–8, turns where RYBKA reported a more than three pawns advantage already at the previous move, and turns involved in repetitions are weeded out.

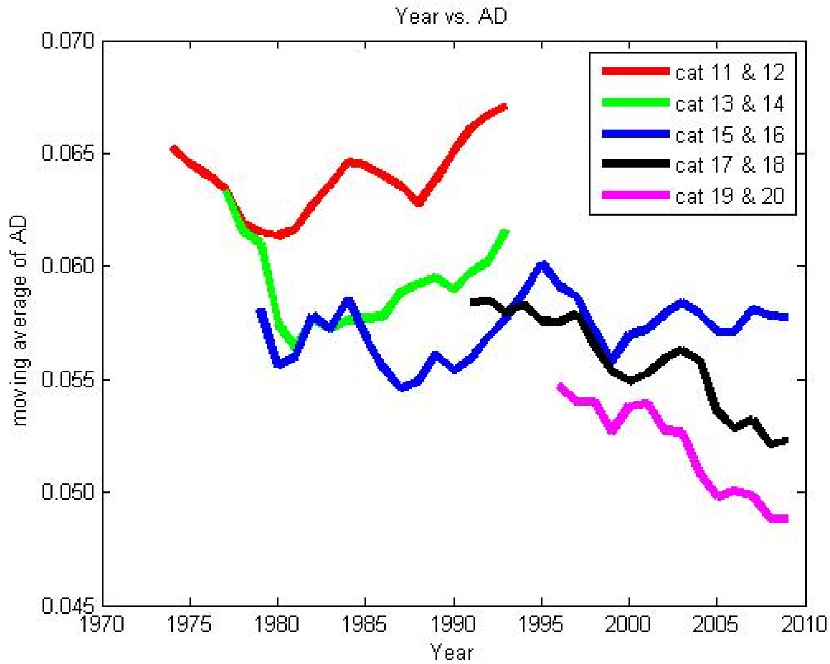
The analysis computations have included all round-robin events of Category 17 or higher, and all Category 11 and higher from 1971 to 1993. The categories are the average rating of players in the event taken in blocks of 25 points; for instance, category 11 means the average rating is between 2500 and 2525, while category 15 means 2600–2625.

For every move that is not equivalent to RYBKA's top move, the "error" is taken as the value of the present position minus the value after the move played. The errors over a game or player-performance or an entire tournament are summed and divided by the number of moves (those not weeded out) to make the "Average Error" (AE) statistic. This omits some refinements of Guid and Bratko's "Average Difference" (AD) statistic [GB06], so we use the separate name.

For large numbers of games, AD or AE seems to give a reasonable measure of playing quality. When aggregated for all tournaments in a span of years, the figures were in fact used to make scale corrections for the in-depth mode presented in the next section. When AE is plotted against the turn number, sharply greater error for turns approaching the standard Move 40 time control is evident; then comes a sharp drop back to previous levels after Move 41. When AE is plotted against the advantage or disadvantage for the player to move, in intervals of 0.10 or 0.05 pawns, a scaling pattern emerges. The AE for advantage 0.51–0.60 is almost double that for near-equality 0.01–0.10, while for -0.51 to -0.60 it is regularly more than double. It would seem strange to conclude that strong masters play only half as well when ahead or behind by half a pawn as even. Rather this seems to be evidence that human players perceive differences in value in proportion to the overall advantage for one side. This yields a log-log kind of scaling, with an additive constant that tests place close to 1, so we used 1.

With all this said, here is a plot of AE for all tournaments by year as a four-year moving average.

[4] That RYBKA versions often report the depth as -2 or -1 in UCI feedback has fueled speculation that the true depth here is 16, while the first author finds it on a par in playing strength with some other prominent programs fixed to depths in the 17–20 range.



The five lines represent categories 11–12 (2500–2549 average rating), 13–14 (2550–2599), 15–16 (2600–2649), 17–18 (2650–2699), and 19–20 (2700–2749). The lowest category has the highest AE and hence appears at the top. To be sure there are variations year-to-year, but the graph allows drawing two clear conclusions: the categories do correspond to different levels of AE, and the lines by-and-large do not slope up to the right as would indicate inflation. Indeed, the downslope of AE for categories above 2650 suggests some deflation since 1990.

Since the AE statistic itself depends on how tactically challenging a game is, and hence does not indicate skill by itself, we need a more intensive mode of analysis in order to judge skill directly.

5 Intrinsic Ratings Over Time

Haworth [Haw03,Haw07] and G. DiFatta and Regan [DHR09,HRD10,RH11] have developed models of fallible decision agents that can be trained on players' games and calibrated to a wide range of skill levels. Their main distinction from [GB06] is the use of Multi-PV analysis to obtain equally authoritative values for all reasonable options, not just the top move(s) and the move played. Thus each move is evaluated in the full context of available options.

Models of this kind function in one direction by taking in game analyses and using statistical fitting to generate values of the skill parameters to indicate the intrinsic level of the games. They function in the other direction by taking pre-set values of the skill parameters and generating a probability distribution of next moves by an agent of that skill profile. The defining equation of the particular model used in [RH11], relating the probability p_i of the i-th alternative move to p_0 for the best move and its difference in value, is

$$\frac{\log(1/p_i)}{\log(1/p_0)} = e^{-(\frac{\delta}{s})^c}, \qquad \text{where} \qquad \delta_i = \int_{v_i}^{v_0} \frac{1}{1+|z|} dz. \tag{3}$$

Here when the value v_0 of the best move and v_i of the i-th move have the same sign, the integral giving the scaled difference simplifies to $|\log(1+v_0) - \log(1+v_i)|$. Note that this employs the empirically-determined scaling law from the last section.

The skill parameters are called s for "sensitivity" and c for "consistency" because s when small can enlarge small differences in value, while c when large sharply cuts down the probability of poor moves. The equation solved directly for p_i becomes

$$p_i = p_0^{\alpha} \qquad \text{where} \qquad \alpha = e^{-(\frac{\delta}{s})^c}.$$

The constraint $\sum_i p_i = 1$ thus determines all values. By fitting these derived probabilities to actual frequencies of move choice in training data, we can find values of s and c corresponding to the training set.

Each Elo century mark 2700,2600,2500,... is represented by the training set comprising all available games under standard time controls in round-robin or small-Swiss (such as no more than 54 players for 9 rounds) in which both players were rated within 10 points of the mark, in the three different time periods 2006–2009, 1991–1994, and 1976–1979. In [RH11], it was observed that the computed values of c stayed within a relatively narrow range, and gave a good linear fit to Elo rating by themselves. Thus it was reasonable to impose that fit and then do a single-parameter regression on s. The "central s, c artery" created this way thus gives a simple linear relation to Elo rating. Then if a player P's games in an event E produce a point (s_P, c_P) on the line when fitted, the corresponding Elo rating can be read right off. Points off the artery can be mapped to the closest point on it by an iterative process.

Here we take a more direct route by computing from any (s, c) a single value that corresponds to an Elo rating. The value is the *expected error per move* on the union of the training sets. We denote it by AE_e, and note that it, the expected number MM_e of matches to the computer's first-listed move, and projected standard deviations for these two quantities, are given by these formulas:

$$\begin{aligned} MM_e &= \textstyle\sum_{t=1}^{T} p_{0.t}, & \sigma_{MM} &= \sqrt{\textstyle\sum_{t=1}^{T} p_{0,t}(1-p_{0,t})} \\ AE_e &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\sum_{i\geq 1} p_{i,t}\delta_{i,t}, & \sigma_{AE} &= \sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\sum_{i\geq 1} p_{i,t}(1-p_{i,t})\delta_{i,t}}. \end{aligned}$$

The first table gives the values of AE_e that were obtained by first fitting the training data for 2006–09, to obtain s, c, then computing the expectation for the union of the training sets. It was found that a smaller set of moves comprising the games of the 2005 and 2007 world championship tournaments and the 2006 world championship match gave identical results to the fourth decimal place.

Elo	2700	2600	2500	2400	2300	2200
AE_e	.0572	.0624	.0689	.0749	.0843	.0883

A simple linear fit then yields the rule to produce the Elo rating for any (s, c), which we call an "Intrinsic Performance Rating" (IPR) when the (s, c) are obtained by analyzing the games of a particular event and player(s).

$$\text{IPR} = 3571 - 15413 \cdot AE_e. \tag{4}$$

This expresses, incidentally, that at least from the vantage of RYBKA 3 run to reported depth 13, perfect play has a rating under 3600. This is reasonable when one considers that if a 2800 player such as Vladimir Kramnik is able to draw one game in fifty, the opponent can never have a higher rating than that.

Using equation (4), we reprise the main table from [RH11], this time with the corresponding Elo ratings from the above formulas. The left-hand side gives the original fits, while the right-hand side corresponds to the "central artery" discussed above.

2006–2009

Elo	s	c	IPR	c_{fit}	s_{fit}	IPR
2700	.078	.503	2690	.513	.080	2698
2600	.092	.523	2611	.506	.089	2589
2500	.092	.491	2510	.499	.093	2528
2400	.098	.483	2422	.492	.100	2435
2300	.108	.475	2293	.485	.111	2304
2200	.123	.490	2213	.478	.120	2192
2100	.134	.486	2099	.471	.130	2072
2000	.139	.454	1909	.464	.143	1922
1900	.159	.474	1834	.457	.153	1802
1800	.146	.442	1785	.450	.149	1801
1700	.153	.439	1707	.443	.155	1712
1600	.165	.431	1561	.436	.168	1565

1991–1994

2700	.079	.487	2630	.513	.084	2659
2600	.092	.533	2639	.506	.087	2609
2500	.098	.500	2482	.499	.092	2537
2400	.101	.484	2396	.492	.103	2406
2300	.116	.480	2237	.485	.117	2248
2200	.122	.477	2169	.478	.122	2173

1976–1979

2600	.094	.543	2647	.506	.087	2609
2500	.094	.512	2559	.499	.091	2547
2400	.099	.479	2397	.492	.103	2406
2300	.121	.502	2277	.485	.116	2257

The entries vary around the Elo century marks, as is to be expected from a linear fit. Some points in the 1600–2100 range are anomalous, and this may owe to various factors pertaining to the quality of the games. The standard deviations are such that only the first 3 digits of the IPR are significant, but the table still supports a conclusion of no overall inflation. Because the fit was done with data from 2006–2009 only, inflation would show up as, for instance, 2600- and 2500-rated players from earlier years having higher IPR's than players with those ratings today.

Further support for our basic contention that today's players deserve their higher ratings comes from historical IPR's. Since the IPR is based only on game analysis and has no functional component from Elo, it extends before the adoption of Elo to the beginning of chess. The next table gives IPR's for some players and events.

Player, player in event, or entire event	IPR
Howard Staunton, versus P. de Saint-Amant	1899
Staunton, all major matches	1940
Adolf Anderssen, London 1851	2004
Anderssen, versus Paul Morphy	2112
Morphy, versus Anderssen	2124
Morphy, 59 most important games overall	2344
Anderssen, 1860 onward	2100
Wilhelm Steinitz, up to 1870	1937
Steinitz, 1871–1882	2320
Steinitz, London 1883	2486
Steinitz, all games versus Zukertort	2352
Steinitz, all games versus Chigorin	2146
Steinitz, all games versus Gunsberg	2495
Steinitz, all games versus Lasker	2334
Johannes Zukertort, all games	2188
Zukertort, London 1883	2445
Zukertort, all games with Steinitz	2199
Emanuel Lasker, all games with Steinitz	2471
St. Petersburg 1896 quadrangular	2390
Cambridge Springs 1904, top 9 vs. each other	2443
St. Petersburg 1914 prelims	2331
St. Petersburg 1914 finals	2575
New York 1927	2580
José Raoul Capablanca at New York, 1927 (not a typo)	2936
Capablanca at AVRO 1938	2681
AVRO 1938 overall	2605
The Hague 1948	2639
Paul Keres at The Hague 1948	2657
Curacao 1962 Candidates' Tournament	2536

Player, player in event, or entire event	IPR
Linares 1993	2521
Anatoly Karpov at Linares 1993	2569
Garry Kasparov at Linares 1993	2813
Linares 1994	2517
Karpov at Linares 1994	2917
Kasparov at Linares 1994	2481
Corus 2006	2735
Corus 2007	2765
Sofia 2006	2744
Sofia 2007	2576
Sofia 2008	2691
Sofia 2009	2705
Nanjing 2010	2747
Shanghai 2010	2831
Bilbao 2010	2906
Moscow Tal Memorial 2010	2693
London Classic 2010	2669

This shows a steady progression in IPR throughout chess history, mirroring the improvement of sporting records in other fields and ascribable to better human health overall, and greater wealth allowing there to be more enthusiasts. This argues against ratings having inflated relative to skill.

It is possible that the model as currently constituted is favoring positional players at the expense of tactical ones. Currently it quantifies only how one reacts to one's own challenges; it may need to quantify the degree of challenge set for the opponent. It may also be overly influenced by plus-scores, as shown in the next section.

6 Distributions of Performances

The final experiment reported here analyzed 220 games from the 2011 Canadian Open, including all from players with FIDE ratings 2400 and above, and all who finished with at least 6/9. The following table shows the IPR's and compares them to Chess Federation of Canada ratings before and after the event, FIDE ratings before, and the tournament performance ratings (TPR's) based on the CFC ratings. The right hand column restricts to games against opponents with FIDE ratings 2400 and above. Players marked with a * had fewer than 100 total analyzed moves against such players, and so are not included in the latter sampling. From this we draw the following tentative conclusions, pending an analysis in progress of the entire tournament:

1. FIDE ratings of Canadian players are deflated relative to apparent skill. This is commonly believed to be due to a lack of playing opportunities in FIDE-rated events.

2. The model gives higher results even than Canadian ratings when all games are included, but lower results when restricting to games among the FIDE 2400+ group.

Name	CanR	post	FIDE	TPR	IPR	Diff	MvGp	IPR+	-FIDE
Arencibia	2537	2556	2476	2745	2722		216	2737	+261
Benjamin	2641	2646	2553	2688	2411	-277	308	2491	- 62
Bluvshtein	2634	2634	2611	2622	2534		267	2462	-149
Bojkov*	2544	2550	2544	2595	2152	-443	64		
Calugar*	2437	2408	2247	2144	2301	+157	0		
Cheng	2500	2514	2385	2661	2732		189	2614	+229
Cummings*	2459	2461	2350	2473	2833	+360	47		
Fedorowicz*	2508	2500	2454	2422	2390		32		
Gerzhoy	2647	2645	2483	2622	2964	+342	192	2934	+451
Golod*	2576	2577	2582	2582	2640		110		
Hambleton	2349	2359	2206	2425	2247	-178	101	1759	-447
Hebert*	2486	2490	2414	2519	2789	+270	79		
Humphreys	2277	2300	2111	2458	2235	-223	180	2021	- 90
Krnan	2470	2486	2390	2651	2694		180	2477	+ 87
Krush	2578	2575	2487	2539	2497		211	2495	+ 8
Macak*	2391	2374	2391	2273	2521	+248	0		
Meszaros*	2409	2399	2418	2278			62		
Mikhalevski	2664	2652	2569	2519	2615		199	2538	- 31
Milicevic*	2400	2397	2288	2352			50		
Mulyar	2422	2421	2410	2412	2635	+223	103	2683	+273
Noritsyn	2597	2594	2425	2563	2393	-170	162	2077	-348
Pechenkin*	2408	2401	2297	2309	2647	+338	11		
Perelshteyn	2532	2543	2534	2650	2629		196	2559	+ 25
Perez Rod'z	2467	2488	2467	2676	2628		116	2617	+150
Plotkin*	2411	2399	2243	2260			52		
Regan*	2422	2408	2409	2268	2525	+257	51		
Rozentalis	2614	2619	2571	2666	2722		222	2742	+171
Sambuev	2739	2723	2528	2571	2676		261	2584	+ 56
Samsonkin	2532	2549	2378	2707	2535	-172	189	2502	+124
Sapozhnikov	2424	2429	2295	2480	2404		165	2231	- 64
Shabalov	2618	2613	2577	2549	2643		131	2814	+237
Thavandiran	2447	2461	2320	2607	2621		128	2427	+107
Wang*	2346	2350	2240	2376	2340		88		
Xu	2149	2185	1875	2522	2151	-371	149	1869	- 6
Yoos*	2439	2426	2373	2289			77		
Zenyuk*	2429	2424	2222	2342	2791	+449	93		
Averages	2486		2392	2495	2551				
21 no *	2516		2412	2587	2557			2459	+ 47
15 *ed	2444		2365	2365	2539				
32 with IPR	2495		2400	2520	2551				

3. There are some strange high outlier IPR's (David Cummings and Iryna Zenyuk, even more than Leonid Gerzhoy), but the averages overall are reasonable—especially when restricted to games within the 2400+ group, all of which were analyzed.

7 Conclusions

In this paper we have shown multiple, separate, and novel pieces of evidence that the population of Elo-rated chess-players has remained stable in the relation of rating to intrinsic skill level, and obeys simple large-scale population laws that make no reference to inflation. We have presented evidence that chess skill has increased steadily throughout history, in line with the increased number of high-rated players. Given this stability in the FIDE system, we can promote the use of our tools in adjusting members of national federations wit their own rating pools to the international scale.

We anticipate further development of the methods in this paper. It is possible that some rating systems being tested as alternatives to Elo in the recent *Kaggle* competitions sponsored by Sonas [SK11] may yield better correspondences to our models.

Acknowledgments Foremost we thank the programmers of the Arena chess GUI for providing full scripting and recording of computer analysis. Tamal Biswas collected data and prepared the chart of AE for tournaments by category. We thank the Universities of Buffalo and Montreal for equipment support, and the programmers of TOGA II and RYBKA 3 for their engines and advice. Finally we thank Canadian Open organizer David Cohen, especially for his help in obtaining game scores from the event.

References

[DHR09] G. DiFatta, G. Haworth, and K. Regan. Skill rating by Bayesian inference. In *Proceedings, 2009 IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09), Nashville, TN, March 30–April 2, 2009*, pages 89–94, 2009.

[dSP61] Derek J. de Solla Price. *Science Since Babylon*. Yale University Press, 1961.

[GB06] M. Guid and I. Bratko. Computer analysis of world chess champions. *ICGA Journal*, 29(2):65–73, 2006.

[Goo94] David Goodstein. The big crunch. In *Proceedings, 48th NCAR Symposium, Portland*, 1994.

[Haw03] G. Haworth. Reference fallible endgame play. *ICGA Journal*, 26:81–91, 2003.

[Haw07] Guy Haworth. Gentlemen, Stop Your Engines! *ICGA Journal*, 30:150–156, 2007.

[HRD10] G. Haworth, K. Regan, and G. DiFatta. Performance and prediction: Bayesian modelling of fallible choice in chess. In *Proceedings, 12th ICGA Conference on Advances in Computer Games, Pamplona, Spain, May 11–13, 2009*, volume 6048 of *Lecture Notes in Computer Science*, pages 99–110. Springer-Verlag, 2010.

[RH11] K. Regan and G. Haworth. Intrinsic chess ratings. In *Proceedings of AAAI 2011, San Francisco*, 2011.

[RK07] V. Rajlich and L. Kaufman. Rybka 3 chess engine, 2007. http://www.rybkachess.com.

[SK11] J. Sonas and Kaggle.com. Chess ratings: Elo versus the Rest of the World. http://www.kaggle.com/c/chess, 2011.

[Son11] J. Sonas. Chessmetrics. http://www.chessmetrics.com, 2011.

[Ver38] Pierre-Francois Verhulst. Notice sur la loi que la population poursuit dans son accroissement, 1838.