Understanding distributions of chess performances

Book or Report Section

Accepted Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

www.reading.ac.uk/centaur

CentAUR

Understanding Distributions of Chess Performances

Kenneth W. Regan[1], Bartlomiej Macieja[2], and Guy M^{c}C. 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 studies the population of chess players and the distribution of their performances measured by 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, and by skill parameters from a model that employs computer analysis keeping a nearly constant relation to Elo rating across that time. 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?

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. We continue work by Haworth, DiFatta, and Regan [1–4] on measuring performance '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 [4] that ratings have remained stable since the 1970's, contrary to the popular wisdom of extensive *rating inflation*. Section 5 extends that work to the Elo scale, while the other sections present independent supporting material. Related previous work [5–7] is discussed below.

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, and
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 [4] 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 statistics in this paper 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 (say) 50 to 100 points higher—as Kasparov did as the 1980's progressed.
(b) Karpov and Korchnoi were supreme strategists whose strategic insight and *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 whether and how they are prepared. Regarding also (b) we find that today's elite make fewer clear mistakes than their forebears. This factor impacts skill apart from strategic depth. The model from [4] used in this paper finds a natural weighting for 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, without adjusting for inflation.
- A version, called AE for "average error," of the "average difference" (AD) statistic used by Guid and Bratko [5] (see also [6, 7]) to compare world championship matches. An important scaling discovery leads to *Scaled Average Error* (SAE). Our work shows that tournaments of a given category have seen fairly constant (S)AE 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 [4].
- 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, with somewhat higher variance. 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 Regan-Haworth model [4] 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. Our non-standard analyses may take the hind legs, but since they all agree, we feelm we understand the elephant. 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 [8], 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, Goodstein [9] 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 Verhulst [10] 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

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

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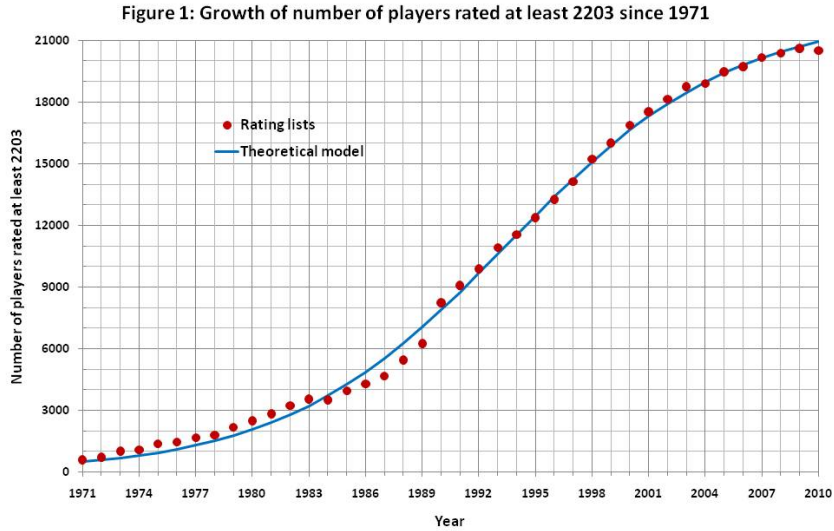
Figure 1 shows the number of 2203+ rated players, and a curve obtained for some particular values of a, b, and N_{max}. Since there are many data points and only three parameters, the fit is striking. This implies that the growth of the number of chess players can be explained without a need to postulate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

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 [11] to fixed reported depth 13 ply[1] in Single-PV mode.

---

[1] 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.

This mode is similar to how Guid and Bratko [5] 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 3.00 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 11 or higher, using all events given categories in the ChessBase Big 2010 database plus The Week In Chess supplements through TWIC 893 12/19/11. 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. Besides including moves 9–12 and using Rybka depth 13 with a $[-3.00, +3.00]$ evaluation range rather than Crafty depth 12 with a $[-2.00, +2.00]$ range, our statistic differs from [5] in not attempting to judge the "complexity" of a position, and in several incidental ways.

For large numbers of games, AD or AE seems to give a reasonable measure of playing quality, beyond relative ranking as shown in [6]. 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. This is reflected in the definition of the scaled difference δ_i in Equation 3 below, since $1/(1 + |z|)$ in the body of a definite integral produces $\ln(1 + |z|)$. This produces *Scaled Average Error* (SAE).

Figure 2 shows AE (called R3 for "raw" and the 3.00 evaluation cutoff) and SAE (SC3), while Figure 3 shows how both figures increase markedly toward the standard Move 40 time control and then level off. For these plots the tournaments were divided into historical "eras" E5 for 1970–1984, E6 for 1985–1999, E7 for 2000–2009, and E8 for 2010–. The tournaments totaled 57,610 games, from which 3,607,107 moves were analyzed (not counting moves 1–8 of each game which were skipped) and over 3.3 million retained within the cutoff. Category 10 and lower tournaments that were also analyzed bring the numbers over 60,000 games and 4.0 million moves with over 3.7 million retained. Almost all work was done on two quad-core Windows PC's with analysis scripted via the Arena GUI v1.99 and v2.01.
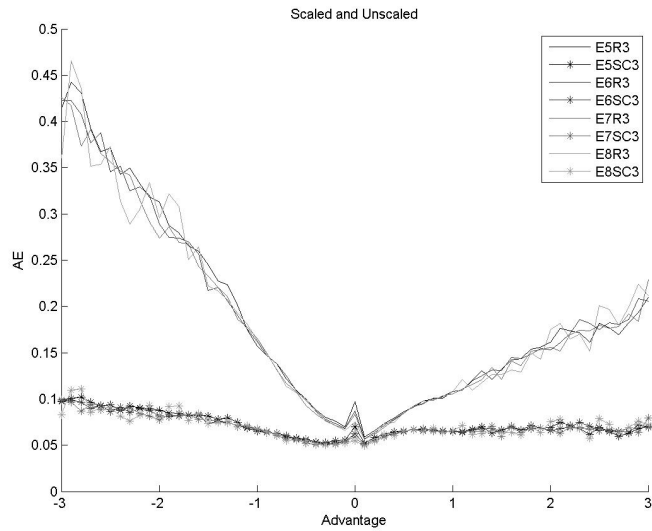
Figure 2: Plot of raw AE vs. advantage for player to move, and flattening to SAE.
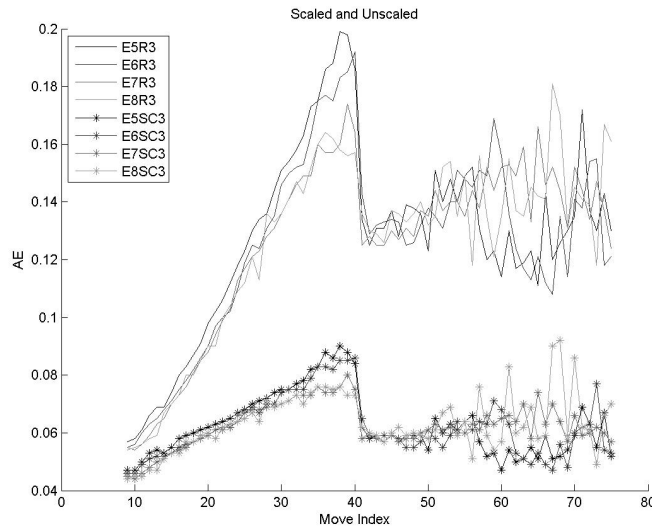
Figure 3: Plot of AE and SAE by turn number.

Figures 4 and 5 below graph SAE for all tournaments by year as a *four-year moving average*, the latter covering moves 17–32 only. The five lines represent categories 11–12 (FIDE Elo 2500–2549 average rating), 13–14 (2550–2599), 15–16 (2600–2649), 17–18 (2650–2699), and 19–20 (2700–2749). There were several category 21 events in 1996–2001, none in 2002–2006, and several 21 and 22 events since 2007; the overall averages of the two groups are plotted as X for 2001 and 2011. The lowest category has the highest SAE and hence appears at the top.
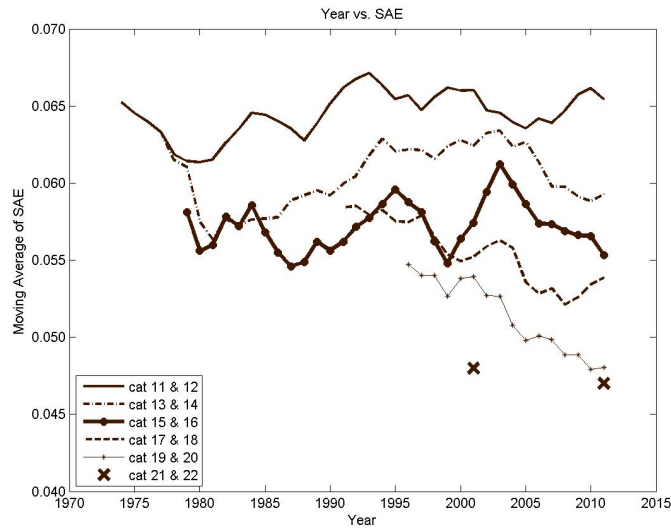
Figure 4: SAE by tournament category, 4-yr. moving avg., 1971–2011.
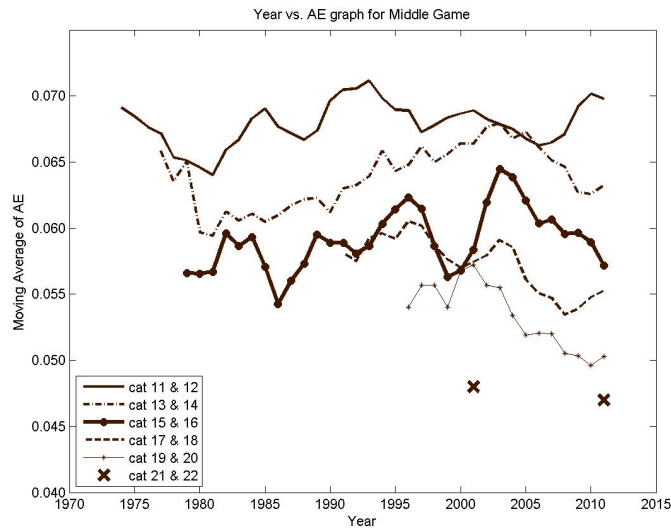
Figure 5: SAE by category for moves 17–32 only, 4-yr. moving avg., 1971–2011.

Despite yearly variations the graphs allow drawing two clear conclusions: the categories do correspond to different levels of SAE, and the lines by-and-large do not slope up to the right as would indicate inflation. Indeed, the downslope of SAE for categories above 2650 suggests some deflation since 1990. Since the SAE statistic 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 [1, 2] and with DiFatta and Regan [3, 12, 4] developed models of fallible decision agents that can be trained on players' games and calibrated to a wide range of skill levels. Their main difference from [5–7] is the use of Multi-PV analysis to obtain uthoritative 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. The paper [6] gives evidence that for relative rankings of players, good results can be obtained even with relatively low search depths, and this is confirmed by [7]. However, we argue that for an intrinsic standard of quality by which to judge possible rating drift, one needs greater depth, the full move context, and a variety of scientific approaches. The papers [3, 12] apply Bayesian analysis to characterize the performance of human players using a spectrum of *reference fallible agents*. The work reported in [4] and this paper uses a method patterned on multinomial Bernoulli trials, and obtains a corresponding spectrum.

The *scaling* of AE was found important for quality of fit, and henceforth AE means SAE. It is important to note that SAE from the last section does not directly carry over to intrinsic ratings in this section, because here we employ the full move analysis of Multi-PV data. They may be expected to correspond in large samples such as all tournaments in a range of years for a given category, but here we are considering smaller samples from a single event or a single player in a single event, and at this stage we are studying those with more intensive data. What we do instead is use statistical fits of parameters called s, c to generate *projections* AE_e for every position, and use the aggregate projected AE_e on a reference suite of positions as our "standard candle" to index to the Elo scale.

We also generate projected standard deviations, and hence projected confidence intervals, for AE_e (and also the first-move match statistic MM_e) as shown below. This in turn yields projected confidence intervals for the intrinsic ratings. Preliminary testing with randomly-generated subsets of the training data suggest that the actual deviations in real-world data are bounded by a factor of 1.15 for the MM statistic and 1.4 for AE, and these are signified by a subscripted a for 'actual' in tables below. The projections represent the ideal case of zero modeling error, so we regard the difference shown by the tests as empirical indication of the present level of modeling error.

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 [4], relating the probability p_i of the i-th alternative move to p_0 for the best move and its difference in value, is

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

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

$$p_i = p_0^\alpha \qquad \text{where} \qquad \alpha = e^{-\left(\frac{\delta}{s}\right)^c}. \tag{4}$$

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 [4], 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.

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:

$$MM_e = \sum_{t=1}^{T} p_{0.t}, \qquad \sigma_{MM_e} = \sqrt{\sum_{t=1}^{T} p_{0,t}(1 - p_{0,t})}$$
$$AE_e = \frac{1}{T} \sum_{t=1}^{T} \sum_{i \geq 1} p_{i,t} \delta_{i,t}, \quad \sigma_{AE_e} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \sum_{i \geq 1} p_{i,t}(1 - p_{i,t}) \delta_{i,t}}. \tag{5}$$

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 R 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, so R was used as the fixed *reference set*.

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

Table 1. Correspondence between Elo rating from 2006–2009 and projected Average Error.

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).

$$IPR = 3571 - 15413 \cdot AE_e. \tag{6}$$

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 (6), we reprise the main table from [4], 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. The middle of the table is our first instance of the following procedure for estimating confidence intervals for the IRP derived from any test set:

1. Do a regression on the test set T to fit s_T, c_T.
2. Use s_T, c_T to project AE_e on the reference set R (not on T), and derive IPR_T via equation (6).
3. Use s_T, c_T on the test set T only to project $\sigma_T = \sigma_{AE_e}$.
4. Output $[\text{IPR}_T - 2\sigma_T, \text{IPR}_T + 2\sigma_T]$ as the proposed "95%" confidence interval.

As noted toward the start of this section, early testing suggests replacing σ_T by $\sigma_a = 1.4\sigma_T$ to get an "actual" 95% confidence interval given the model as it stands. Hence we show both ranges.

In this case, the test sets T are the training sets themselves for the Elo century points in three different four-year intervals. These give the results in Table 2.

				2006–2009					
Elo	s	c	IPR	$2\sigma_e$ range	$2\sigma_a$ range	#moves	c_{fit}	s_{fit}	IPR_{fit}
2700	.078	.502	2690	2648–2731	2632–2748	7,032	.513	.080	2698
2600	.092	.523	2611	2570–2652	2553–2668	7,807	.506	.089	2589
2500	.092	.491	2510	2480–2541	2468–2553	16,773	.499	.093	2528
2400	.098	.483	2422	2393–2452	2381–2464	20,277	.492	.100	2435
2300	.108	.475	2293	2257–2328	2243–2342	17,632	.485	.111	2304
2200	.123	.490	2213	2170–2257	2153–2274	11,386	.478	.120	2192
2100	.134	.486	2099	2048–2150	2028–2170	9,728	.471	.130	2072
2000	.139	.454	1909	1853–1966	1830–1989	9,471	464	.143	1922
1900	.159	.474	1834	1790–1878	1769–1893	16,195	.457	.153	1802
1800	.146	.442	1785	1741–1830	1723–1848	15,930	.450	.149	1801
1700	.153	.439	1707	1642–1772	1616–1798	8,429	.443	.155	1712
1600	.165	.431	1561	1496–1625	1470–1651	9,050	.436	.168	1565
				1991–1994					
2700	.079	.487	2630	2576–2683	2555–2704	4,954	.513	.084	2659
2600	.092	.533	2639	2608–2670	2596–2682	13,425	.506	.087	2609
2500	.098	.500	2482	2453–2512	2441–2524	18,124	.499	.092	2537
2400	.101	.484	2396	2365–2426	2353–2438	19,968	.492	.103	2406
2300	.116	.480	2237	2204–2270	2191–2284	20,717	.485	.117	2248
2200	.122	.477	2169	2136–2202	2123–2215	21,637	.478	.122	2173
				1976–1979					
2600	.094	.543	2647	2615–2678	2602–2691	11,457	.506	.087	2609
2500	.094	.512	2559	2524–2594	2509–2609	11,220	.499	.091	2547
2400	.099	.479	2397	2363–2431	2350–2444	16,635	.492	.103	2406
2300	.121	.502	2277	2240–2313	2226–2328	15,284	.485	.116	2257

Table 2. Elo correspondence in three four-year intervals.

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. Only the Elo 2200 through 2700 data for 2006–2009 were used in the linear fit for the ratings. Of course there is error from the regression, but we do not know whether it adds to or mitigates the estimates σ_{AE_e} of placement of the linear regression points. For uniformity with later performance testing, we show only the latter error here. Despite these elements of uncertainty, 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 positions comes from IPR's of entire tournaments. Table 3 shows all twenty Category 21 or higher round-robin tournaments ever played under standard time controls, while Table 4 shows some others for comparison.

Event	cat: Elo	IPR	$2\sigma_e$ range	$2\sigma_a$ range	IPR-Elo	#moves
Las Palmas 1996	21: 2756	2697	2612–2781	2579–2815	-59	1,760
Linares 1998	21: 2752	2715	2651–2780	2625–2805	-37	2,717
Linares 2000	21: 2751	2728	2645–2810	2612–2843	-23	1,636
Dortmund 2001	21: 2755	2752	2760–2834	2637–2866	-3	1,593
Mexico 2007	21: 2751	2708	2647–2769	2623–2793	-43	3,213
Morelia-Linares 2008	21: 2755	2855	2808–2903	2789–2922	+100	3,453
Nanjing 2008	21: 2751	2766	2691–2842	2660–2873	+15	1,936
Bilbao GSF 2008	21: 2768	2801	2731–2872	2702–2900	+33	2,013
Linares 2009	21: 2755	2750	2696–2803	2675–2825	-5	3,830
Sofia M-Tel 2009	21: 2754	2711	2626–2795	2592–2829	-51	1,937
Nanjing 2009	21: 2763	2715	2644–2785	2616–2814	-48	2,192
Moscow Tal Mem. 2009	21: 2763	2731	2663–2800	2635–2827	-32	2,706
Linares 2010	21: 2757	2681	2607–2756	2577–2786	-76	2,135
Nanjing 2010	21: 2766	2748	2674–2821	2645–2850	-18	1,988
Shanghai 2010	21: 2759	2829	2727–2931	2686–2972	+70	920
Bilbao 2010	22: 2789	2904	2822–2987	2788–3020	+115	1,060
Moscow Tal Mem. 2010	21: 2757	2690	2629–2750	2604–2775	-67	3,493
Bazna 2011	21: 2757	2750	2675–2825	2645–2855	-7	1,885
Sao Paulo/Bilbao 2011	22: 2780	2626	2539–2713	2504–2748	-154	1,998
Moscow Tal Mem. 2011	22: 2776	2807	2755–2860	2734–2881	+31	3,401
Averages	21: 2761	2748			-13	2,293
Weighted by moves	21: 2760	2745			-15.6	
Aggregate run, all moves	21: 2760	2744	2729–2760	2722–2766	-16	45,870

Table 3. Intrinsic Ratings of Category 21 and higher standard tournaments.

The IPR's are on-balance below the tournament average ratings, but the latter's aggregate is just within the narrower confidence interval of the aggregate IPR. The regressions are *not* linear, so the parity of the aggregate run with the weighted average is notable. The comparison events are selective but still show no inflationary trend.

Event	cat: Elo	IPR	$2\sigma_e$ range	$2\sigma_a$ range	IPR-Elo	#moves
Montreal 1979	15: 2622	2588	2534–2642	2513–2663	-34	4,732
Linares 1993	18: 2676	2522	2469–2574	2449–2595	-154	6,129
Linares 1994	18: 2685	2517	2461–2574	2438–2596	-168	5,536
Dortmund 1995	17: 2657	2680	2615–2744	2589–2770	+23	2,459
Dortmund 1996	18: 2676	2593	2518–2667	2489–2697	-83	2,796
Dortmund 1997	18: 2699	2639	2569–2709	2541–2737	-60	2,583
Dortmund 1998	18: 2699	2655	2579–2732	2548–2762	-44	2,284
Dortmund 1999	19: 2705	2749	2655–2844	2617–2882	+44	1,364
Sarajevo 1999	19: 2703	2664	2592–2737	2563–2766	+19	2,755
San Luis 2005	20: 2738	2657	2597–2716	2574–2740	-81	3,694
Corus 2006	19: 2715	2736	2693–2779	2676–2797	+21	5,800
Sofia M-Tel 2006	20: 2744	2744	2678–2810	2651–2836	0	2,197
Corus 2007	19: 2717	2763	2716–2811	2697–2829	+46	5,095
Sofia M-Tel 2007	19: 2725	2576	2482–2670	2445–2708	-149	2,184
Sofia M-Tel 2008	20: 2737	2690	2605–2775	2571–2809	-47	1,869
London Classic 2010	20: 2725	2668	2594–2742	2565–2771	-57	2,312

Table 4. Some other events, for comparison to Table 3.

6 Distributions of Performances

Our final experiment analyzed all 624 available games from 647 played at the 2011 Canadian Open, including all by players with FIDE ratings 2400 and above, which form an unbiased sample. Table 5 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 final two columns are the confidence intervals for the IPR alone. The final rows summarize the sample, the whole event (152 players minus 3 early withdrawals leaving 149), and the whole event weighted by number of games played and number of analyzed moves. The bottom-right restricts to the 115 players who had FIDE ratings before the event. We conclude:

1. The IPR's have similar overall average to the Canadian ratings, especially under weighting by games or moves.
2. 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.
3. The IPR's have higher deviations from their own mean than the TPR's.
4. The IPR's have large deviation, and yet several TPR's fall outside even the 2.8-sigma range. This may constrain the usefulness of the IPR as an estimator of the TPR.

7 Conclusions

We have shown multiple, separate, and novel pieces of evidence that the Elo system employed by FIDE has remained stable in relation to intrinsic skill level. We have shown that the population of master-level players closely fits a model that has an important scientific pedigree, under conditions of no inflation. We have shown that ratings

Name	Can R	FIDE R	TPR	IPR	IPR-TPR	$2\sigma_e$ range	$2\sigma_a$ range	#moves
Arencibia	2537	2476	2745	2723	-22	2491–2956	2398–3049	273
Benjamin	2641	2553	2688	2412	-276	2196–2629	2110–2715	373
Bluvshtein	2634	2611	2622	2533	-89	2323–2744	2239–2828	316
Bojkov	2544	2544	2595	2154	-441	1765–2543	1610–2698	219
Calugar	2437	2247	2144	2301	+157	2091–2512	2007–2596	327
Cheng	2500	2385	2661	2728	+67	2502–2954	2411–3044	297
Cummings	2459	2350	2473	2833	+360	2683–2983	2623–3043	322
Fedorowicz	2508	2454	2422	2390	-32	2088–2692	1967–2813	199
Gerzhoy	2647	2483	2622	2963	+341	2802–3124	2738–3189	211
Golod	2576	2582	2582	2638	+56	2376–2899	2272–3003	218
Hebert	2486	2414	2519	2789	+270	2598–2979	2522–3055	285
Krnan	2470	2390	2651	2694	+43	2488–2900	2405–2982	266
Krush	2578	2487	2539	2497	-42	2217–2717	2189–2805	316
Meszaros	2409	2418	2278	2413	+133	2219–2607	2141–2684	337
Mikhalevski	2664	2569	2519	2616	+96	2412–2820	2330–2902	248
Milicevic	2400	2288	2352	2113	-240	1799–2426	1674–2552	214
Mulyar	2422	2410	2412	2636	+224	2483–2788	2422–2849	378
Noritsyn	2597	2425	2563	2394	-171	2166–2621	2075–2713	286
Pechenkin	2408	2297	2309	2648	+339	2439–2857	2355–2940	311
Perelshteyn	2532	2534	2650	2629	-21	2425–2833	2343–2915	258
Perez Rod'z	2467	2467	2676	2627	-49	2321–2933	2198–3056	195
Plotkin	2411	2243	2260	2715	+455	2570–2861	2512–2919	330
Regan	2422	2409	2268	2525	+257	2323–2728	2242–2809	356
Rozentalis	2614	2571	2666	2721	+55	2528–2913	2452–2990	291
Sambuev	2739	2528	2571	2677	+106	2499–2855	2428–2926	400
Samsonkin	2532	2378	2707	2535	-172	2267–2802	2159–2910	233
Sapozhnikov	2424	2295	2480	2404	-76	2203–2605	2122–2685	341
Shabalov	2618	2577	2549	2639	+90	2417–2861	2328–2590	262
Thavandiran	2447	2320	2607	2622	+15	2360–2884	2255–2989	254
Yoos	2439	2373	2289	1939	-350	1607–2271	1474–2404	268
Zenyuk	2429	2222	2342	2790	+448	2606–2975	2532–3049	229
Averages	2516	2429	2508	2558	+50			
Std. Dev.	92		157	218				
Whole event:	149					Restricted to FIDE-rated players: 115		
Average	2144		2142	2117		2203	2211	2139
Std. Dev.	258		261	379		345	229	220
Wtd. avgs.						IPR	CanR	FIDE R
By games	2156		2154	2134		2219	2221	2147
By moves	2173		2172	2161		2242	2236	2161

Table 5. Comparison of FIDE and CFC ratings, TPR's, and IPR's for 2011 Canadian Open

as reflected in tournament categories have no overall inflationary trend relative to two measures of skill, the simple AE statistic on a large scale embracing (nearly) all tournaments with at least 2500 average rating since 1971, and the more-intensive IPR statistic for some tournaments. We have also furthered the correspondence between Elo century marks and our model's fitted skill parameters shown in [4]. The IPR statistic is the weightiest evidence, but it is important that the other factors give it independent support. Given this stability in the FIDE system, we can promote the use of our tools in adjusting members of national federations with 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 [13, 14] may yield better correspondences to our models.

References

1. Haworth, G.: Reference fallible endgame play. ICGA Journal **26** (2003) 81–91
2. Haworth, G.: Gentlemen, Stop Your Engines! ICGA Journal **30** (2007) 150–156
3. DiFatta, G., Haworth, G., Regan, K.: 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. (2009) 89–94
4. Regan, K., Haworth, G.: Intrinsic chess ratings. In: Proceedings of AAAI 2011, San Francisco. (2011)
5. Guid, M., Bratko, I.: Computer analysis of world chess champions. ICGA Journal **29** (2006) 65–73
6. Guid, M., Pérez, A., Bratko, I.: How trustworthy is Crafty's analysis of world chess champions? ICGA Journal **31** (2008) 131–144
7. Guid, M., Bratko, I.: Using heuristic-search based engines for estimating human skill at chess. ICGA Journal **34** (2011) 71–81
8. de Solla Price, D.J.: Science Since Babylon. Yale University Press (1961)
9. Goodstein, D.: The big crunch. In: Proceedings, 48th NCAR Symposium, Portland. (1994)
10. Verhulst, P.F.: Notice sur la loi que la population poursuit dans son accroissement (1838)
11. Rajlich, V., Kaufman, L.: Rybka 3 chess engine (2007) http://www.rybkachess.com.
12. Haworth, G., Regan, K., DiFatta, G.: 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., Springer-Verlag (2010) 99–110
13. Sonas, J.: Chessmetrics. http://www.chessmetrics.com (2011)
14. Sonas, J., Kaggle.com: Chess ratings: Elo versus the Rest of the World. http://www.kaggle.com/c/chess (2011)