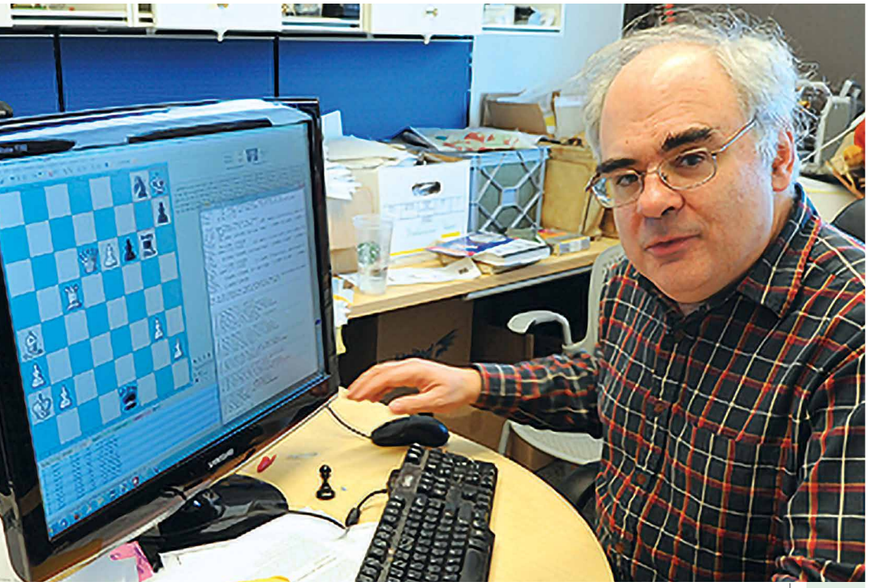


Kenneth Regan is an American professor, chess player, statistician, and computer scientist. At the age of 13 he obtained the USCF Master title, and at age 22 he became an IM. Regan is a professor in the Department of Computer Science and Engineering at the University at Buffalo. He is an expert in anti-cheating in chess, and was involved in investigating known cheaters such as Sebastien Feller, Borislav Ivanov, and Igors Rausis. He commented during the Carlsen–Niemann controversy that his analysis found no evidence of Niemann cheating over the board, though he still largely endorsed Chess.com's report which claimed that Niemann had cheated in numerous online games.

CHANGES IN FIDE CHESS RATINGS IN 2024

Will it become fair enough?

The governing body for chess, FIDE (International Chess Federation), announced changes in its rating system that will take effect on March 1, 2024 – the crucial one being to lift the floor from 1000 to 1400. Are these changes sufficient in order to have a correct system of measuring a player's strength?

By Dusan Krunic

After the COVID pandemic, FIDE applied changes in the rating system back in 2022, announcing that more short-term adjustments might be required. Now, it seems like the time has come for these adjustments. Can you briefly explain the grounds upon which FIDE has decided to apply changes in the rating system?

Have you ever realized that the chess rating system can be used for betting – on other sports and games besides chess? The FiveThirtyEight website started doing exactly that for the major U.S. sports, and others have taken up the mantle such as Neil Paine at his Substack.

Paine currently gives my hometown Buffalo Bills an Elo rating of 1658. This doesn't mean that they are B-player level at football. When the Bills hosted the New England Patriots on New Year's Eve, the Patriots were rated 1398. The difference being 260 meant that the Bills' chances were like yours against an opponent 260 Elo lower. Using either FIDE's official Elo probability table or the table at the TPP sports betting page (https://www.thepunterspage.com/elo-ratings-in-betting), you can see that gave the Bills about an 82% chance of winning. That they were the home team upped the odds a little, much like if you had White. In fact, the Bills won a close game 27–21. They didn't cover the points spread, but a win is a win.

The idea that a difference of ratings gives odds of winning is the same for any sport or game that implements Elo ratings. Tennis and rugby use it; the games of Go and backgammon use it; League of Legends and other online games use it... FIFA adopted a form of Elo in 2018 – well, it doubtless helped that FIDE head Arkady Dvorkovich was the Chair of Organizers of the 2018 FIFA World Cup in Russia. To summarize:

1. Elo ratings give a shorthand for figuring a player or team's chance of winning – so you can set betting odds accordingly.
2. Only the difference in ratings between two opponents matters.
3. The effect of the difference should be the same across the rating spectrum. That is, a 1658 player versus a 1398 player should have the same odds as a 2658 player versus a 2398 player.

In order for these properties to hold, the system needs to regulate itself so that ratings really reflect current skill in a uniform way. The rules

for figuring your new rating after a game or tournament have generally done this remarkably well. But just like a car engine can overheat or leak oil, it is possible for the system to get out of whack. This happened to FIDE ratings of sub-2000 players even before the pandemic cut the fuel lines.

What is the "magic" behind 2000 FIDE Elo?

Jeff Sonas has been one of FIDE's main advisors on ratings for over two decades. Together with Mark Glickman, who has also been a main go-to-guy for the USCF, he created the Universal Rating System (http://universalrating.com) for common rating of blitz, rapid, and classical chess. His ChessMetrics site (http://chessmetrics.com/cm) retro-computes Elo ratings all the way back in chess history. Over years of watching how well the forecast odds play out in game results, he saw the system going to the 'dogs' in the 2010s and decided last year it needed overhaul.

Going to the *underdogs*, that is. Instead of 1398-facing-1658 being under 20% game points expectation, the underdogs were scoring over 30%. If betting on amateur chess games were "a thing," then bookies relying on Elo ratings would go bust. Sonas gives reasons including giving initial/provisional ratings too loosely and amateurs playing too few FIDE-rated events compared to non-FIDE-rated, but we don't have to know the why to observe the what. Here is Sonas' table of how much the favorites have been dogging it since in-person chess started coming back in 2021: *See image 01*

Only the bottom-right grid of matchups where both players are above 2000 is relatively free of skew. The basic math is that below 2000, differences in ratings are overstated. The ratings have become too spaced out. Sonas's prescription is blunt: cut the space by shoehorning everyone from the present FIDE minimum 1000 rating to 2000 into the range 1400-to-2000. As Yasser Seirawan is fond of saying, it's time to do "student body right!"

You've done a lot of work in following the progress of young players in the past decade or so. What were your findings?

I was blissfully unaware that the skew had become so bad even before the pandemic. I have been busy correcting a different skew caused by the brains of improving youngsters expanding by 500–1,000 Elo points while their posted ratings were frozen low. To use myself as a comparison point: after I started playing USCF-rated tournaments in August 1970 my initial ratings came out about 1400 by early 1971. Then in two years I zoomed to 2200+ in 1973. If the pandemic had happened then, so that ratings were frozen in those years, a master expecting an easy lunch against "a 1400" would have been in for quite a shock.

> The effect of the difference should be the same across the rating spectrum. That is, a 1658 player versus a 1398 player should have the same odds as a 2658 player versus a 2398 player.

> Using curves for relationships that the theory of the Elo system says really should be linear is a mathematical original sin.

But as I expanded and refined my own work on chess cheating over the years 2010 to 2019, I unknowingly dealt with the skew in a different way. My work builds on raw metrics of concordance to chess programs: the "T1-match" of how frequently a player makes the same move a computer recommends – or a variant I call "EV-match" counting other moves if they have equal value to the computer's choice – and my scaled-down version of "average centipawn loss" caused by moves that the engine judges inferior. At the time of my stem paper with the late Guy Haworth of Britain's University of Reading in 2011, we had large enough data only on players in the range 1600 to 2700 (FIDE ratings), and the data 1600-to-2000 was recognizably wonky. Allowing for that wonk, I observed a nearly-perfect linear relationship of these metrics to Elo rating – as I expected given property 3 above.

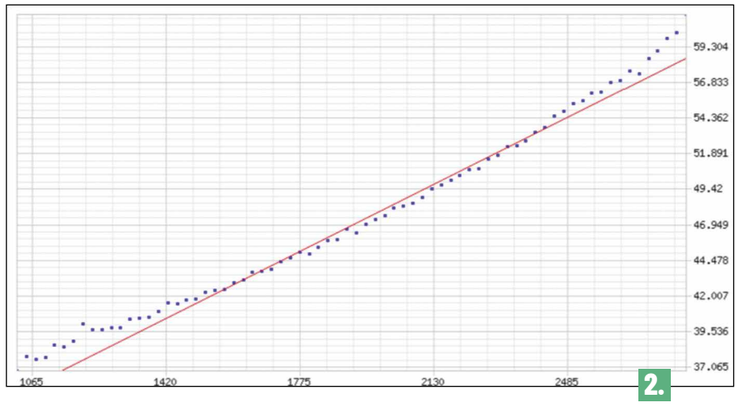
As both databases and amateur play expanded greatly, and as multiple elite players topped 2800, I got voluminous data over the whole 1000-to-2800+ spectrum of ratings. And to my chagrin, the linear relationship recognizably broke down. When I graphed each metric individually, it showed a curve that no longer fitted a line – like this for T1-match to the Stockfish 9 program over the years up to early 2019: *See image 02 (next page)*

I incorporated the curves into both my quick-check "screening formulas" and my full detection system in new editions at the end of 2019 – which are the system I've deployed throughout the pandemic. Once I combined this with my estimation of young players' true skill curves – which I already quantified precisely by the end of 2020 – my system stayed accurate despite the bleeding that Sonas was observing. Or rather, my curves were already acting like a tourniquet to compress the rating space. But using curves for relationships that the theory of the Elo system says really should be linear is a mathematical original sin.

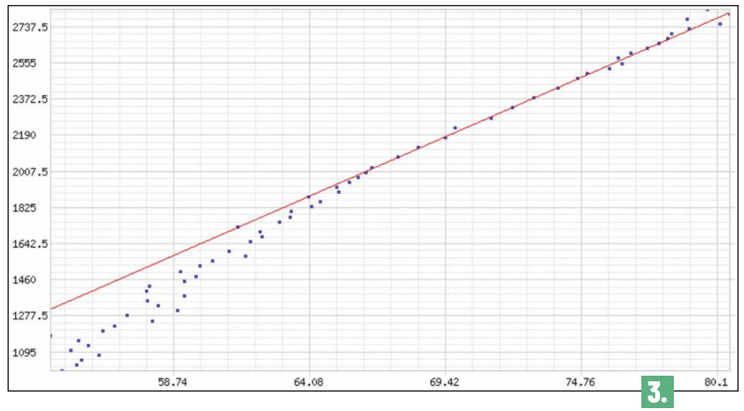
1.

2021-2023	opponent / player	vs. 1000-99	vs. 1100-99	vs. 1200-99	vs. 1300-99	vs. 1400-99	vs. 1500-99	vs. 1600-99	vs. 1700-99	vs. 1800-99	vs. 1900-99	vs. 2000-99	vs. 2100-99	vs. 2200-99	vs. 2300-99	vs. 2400-99	vs. 2500+
+7.4% in 47,760 games	**1000-99**		+2%	+7%	+9%	+10%	+10%	+10%	+9%	+8%	+5%	+5%	+4%	+2%	=0%	=0%	=0%
+7.8% in 150,873 games	**1100-99**	-2%		+4%	+7%	+9%	+12%	+12%	+10%	+9%	+7%	+6%	+5%	+4%	+1%	+1%	=0%
+6.8% in 205,982 games	**1200-99**	-7%	-4%		+5%	+9%	+12%	+14%	+13%	+12%	+9%	+8%	+6%	+4%	+2%	+1%	=0%
+5.3% in 243,185 games	**1300-99**	-9%	-7%	-5%		+5%	+11%	+14%	+15%	+14%	+12%	+10%	+9%	+7%	+4%	+3%	+3%
+3.7% in 270,932 games	**1400-99**	-10%	-9%	-9%	-5%		+6%	+11%	+14%	+15%	+14%	+12%	+9%	+7%	+5%	+2%	+3%
+1.6% in 291,021 games	**1500-99**	-10%	-12%	-12%	-11%	-6%		+6%	+11%	+14%	+14%	+13%	+10%	+9%	+6%	+4%	+0%
+0.1% in 303,892 games	**1600-99**	-10%	-12%	-14%	-14%	-11%	-6%		+6%	+10%	+13%	+14%	+13%	+10%	+7%	+5%	+3%
-1.3% in 318,533 games	**1700-99**	-9%	-10%	-13%	-15%	-14%	-11%	-6%		+5%	+10%	+12%	+11%	+11%	+9%	+6%	+3%
-2.3% in 322,168 games	**1800-99**	-8%	-9%	-12%	-14%	-15%	-14%	-10%	-5%		+5%	+9%	+10%	+10%	+10%	+9%	+3%
-3.2% in 309,028 games	**1900-99**	-5%	-7%	-9%	-12%	-14%	-14%	-13%	-10%	-5%		+5%	+8%	+9%	+9%	+7%	+4%
-3.8% in 275,731 games	**2000-99**	-5%	-6%	-8%	-10%	-12%	-13%	-14%	-12%	-9%	-5%		+4%	+7%	+7%	+7%	+4%
-3.8% in 229,272 games	**2100-99**	-4%	-5%	-6%	-9%	-9%	-10%	-13%	-11%	-10%	-8%	-4%		+4%	+6%	+6%	+5%
-3.9% in 174,867 games	**2200-99**	-2%	-4%	-4%	-7%	-7%	-9%	-10%	-11%	-10%	-9%	-7%	-4%		+4%	+4%	+3%
-3.7% in 128,497 games	**2300-99**	=0%	-1%	-2%	-4%	-5%	-6%	-7%	-9%	-10%	-9%	-7%	-6%	-4%		+3%	+2%
-3.2% in 97,087 games	**2400-99**	=0%	-1%	-1%	-3%	-2%	-4%	-5%	-6%	-9%	-7%	-7%	-6%	-4%	-3%		+2%
-2.5% in 62,886 games	**2500+**	=0%	=0%	=0%	-3%	-3%	-0%	-3%	-3%	-3%	-4%	-4%	-5%	-3%	-2%	-2%	

How do these results of your work match with FIDE (Sonas') suggestions?
News of Sonas' proposal led me to reassess my internal calibration. It occurred to me to graph a hybrid version of all my metrics. That is, credit the player for making one of the computer's top-3 moves, so long as that move had no more than 50 centipawn (figuratively, half a pawn) inferiority. Again, I used only the years 2010—2019, to skirt the pandemic's disruption of ratings. The results – after putting ratings on the vertical axis and using the latest Stockfish version 16 — look like this on Andrew Que's online polynomial regression calculator:
See image 03



2.



3.

This is a kinked line rather than a curve. The kink occurs right near 2000. While Sonas's choice of 2000 as his upper limit may have seemed an arbitrary round number, here it is organic. The way to undo a kinked line is exactly the kind of "compression" he proposes.

The new range should start where the red line is above the bottom-most dot, one line in from the left edge, which is for 1000 rating. This starting point is about Elo 1335. The difference from his proposed starting point of 1400 is minimal — certainly compared to an alternative idea of raising the floor only to 1200.

These three grounds: kinked line not curve, upper limit right near 2000, and lower limit not far from 1400, constitute an independent confirmation of what Sonas has observed and wants to correct. It is independent, because I am using direct quality metrics rather than results of games. Because the metrics are raw rather than from my model calibration, they are also independent of the curved figuring that I apply in my main work. Thus, I put myself on record as supporting his proposal as written.

If we take a look at the top of the world rating list back in March 2019 there were three players above 2800, 4th at 2797 and 5th at 2790. At present, it's only Carlsen and Caruana above 2800 and Nakamura 3rd with 2788. Is it correct to say that the rating deflation exists even among the world's elite?
This is the kind of knock-on effect to expect from the root cause of established players facing improving youngsters whose skill is hundreds of points above their official ratings not yet caught up from being frozen during the pandemic. The adult comes out 5 or so lower than a correct rating would have given whether the game result is win, lose, or draw.

Measuring the deflation directly, via my Intrinsic Performance Rating (IPR) measurements which are grounded in games played from 2010 to 2019, is complicated by several factors. One is that the guesstimated 25—50 Elo effect at elite levels is close to natural uncertainty in ratings themselves. I've intended to show this for the well-controlled dataset of players in the Women's Grand Prix, but did not get time in a busy fall term.

> What it will mean for most amateur players is: more rating points!

Ratings vary in the range from 1400 to, say, 2800. Apart from money, how can strong players get motivated to play more frequently in the Swiss open events? We had a recent example when a 2700+ player withdrew from a tournament after scoring 4/5, judging that his score at that moment will inevitably lead to losing rating points even if he wins all his remaining games.
This was a special situation of the race for the rating-based Candidates berth, and GM Dominguez was paired with several underrated youngsters. Whether elite players would lose rating in Opens compared to team and round-robin events pre-2020 is another idea for study. But I certainly agree that the current status is a disincentive.

With more and more new players joining the game and becoming a part of the rating system, is there a way to prevent the rating deflation? What could have been noticed, particularly after the pandemic, is the situation with young players coming from India — can we possibly have a rating system in which one will have no doubts whether a 12-year-old rated 1600 is certainly weaker than a 12-year-old rated 2300, for example?
Not just India. My rating adjustment formula has worked well around the globe. If anything, the fact of chess going online during the pandemic homogenized the means of improvement, which is what my metrics measure based on directly play, not game outcomes. The rating lag is magnified in India because of isolation and greater proportion of young players. Alas, I have game scores for only a scant few events, so little way to tell directly. I do feel that absent a comprehensive reset of all ratings — based on my measurements where available and interpolation where not — the problems will be significant for a long time. The proposal by Jeff Sonas will fix only some of it.

> North America needs action to make more tournaments be FIDE-rated.

Following the previous example, is there a mathematical model for a minimum number of rated games to properly match the playing strength of a player who had just recently got his/her first rating?
Sonas references the low sample sizes of entry-level players as a cause of the skew below 2000. Modeling this has been more the province of Mark Glickman — it is a basic element of his Glicko system.

Glicko has a second player-specific parameter, called RD for "rating deviation," which quantifies the uncertainty in the player's rating. New players have higher RD. Unlike the K-factor of the standard Elo system, RD is involved in prediction as well as the update rule. One basic effect of Glickman's formulas is that more uncertainty favors the underdog. This is borne out in practice and explains why it is actually correct for Sonas's simulations after his fix to retain a slight plus for the lower-rated player. I explained this effect on my blog post back in 2018 and also about Glickman's work sleuthing which Beatles songs were mostly John versus mostly Paul.

What do you see as main effects of the new FIDE rating regulations?
I do not have comparable large data from USCF-rated games over the same years 2010—2019, and I have not been made aware of any similar prior discussion about USCF ratings. Going ahead, however, there will need to be some adjustment in order to have better correspondence to FIDE ratings. A main reason for FIDE's delay to March 2024 was to give national federations more time for this.

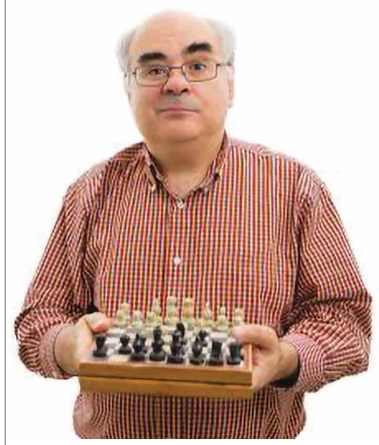
What it will mean for most amateur players is: more rating points! Now this isn't like getting free money. The basic observation is that most amateur players have long been underrated already — at least on the FIDE scale. This should go all the more for those of you who have played lots of USCF (or other national) events that are not FIDE-rated, but only the occasional FIDE event. You and your worldwide comperes — those of you who stay active — have evidently improved apart from the watch of FIDE play.

What it means for me is revamping all my modeling to be once again pristinely linear. Just like a person getting new glasses, I hope to notice sharper contours in my results.

The existence of national ratings in countries like the USA, the UK, Germany, Canada, to name a few, show significant discrepancies in players' ratings — national compared to international. Still, national ratings mean nothing when it comes to obtaining international titles. Do these dual ratings produce confusion, or more clarity? Would it be possible to have only one rating system?
As long as it is not practical for every local tournament to meet FIDE's specifications, there will be separate rating domains. North America needs action to make more tournaments be FIDE-rated. In Britain I observe that the local ratings have caught up much faster from the pandemic — they come even closer to targets than what I get with my adjustments to FIDE ratings.

> I think what we need is a one-time reset of ratings under some neutral criteria, and then it will be OK to operate as in the past.

When you mention other sports, like tennis for example, the probability of winning may be well-implemented with the tables you refer to. In chess it is literally binary (1 and 0, if we omit the draw), whereas in tennis you receive a different amount of points for reaching different phases of the tournament (ATP list is created based on these points and not the pure result of a single match). Is it the difference in the scoring system that may be an obstacle in chess compared to tennis?
How you update rating points and how they predict results are separate matters. Ratings predict a percentage of points under any scoring system. In chess, the 260-point difference in my article really says the chances **p** of a win and **q** of a draw are such that **p + 0.5q = 0.817**, without necessarily projecting **p** and **q** by themselves. Updating with extra points for, say, reaching the semis of a knockout event may be analogous to how the USCF has tacked on bonus points to a large gain. What Sonas and Glickman and other FIDE and USCF people do is frame and tweak simple update rules to optimize how well they (would have) predicted recent past results. Sonas conveys this in his proposal.



In your opinion, what would be other huge changes in the direction of establishing an ideal rating system?
Under most conditions the basic rating system works well. My earlier work shows that FIDE ratings remained remarkably stable up through ten years ago — no big inflation as commonly alleged. This is relative to my IPR metric. The improvements found in "Kaggle" competitions that were held before 2020 mostly use information beyond a player's current rating and the opponents, and game results of the present tournament to do their updates. They lose the appeal of simply updating one's rating game-by-game, which made GM Firouzja's quest in December gripping to follow. The current troubles happened first because the population of players lost stability at the entry points, and then the pandemic upended everything.

So, I think what we need is a one-time reset of ratings under some neutral criteria, and then it will be OK to operate as in the past.