Compression and Calculation Improvements: Final Report

-- by Jeff Sonas (11 January 2024)

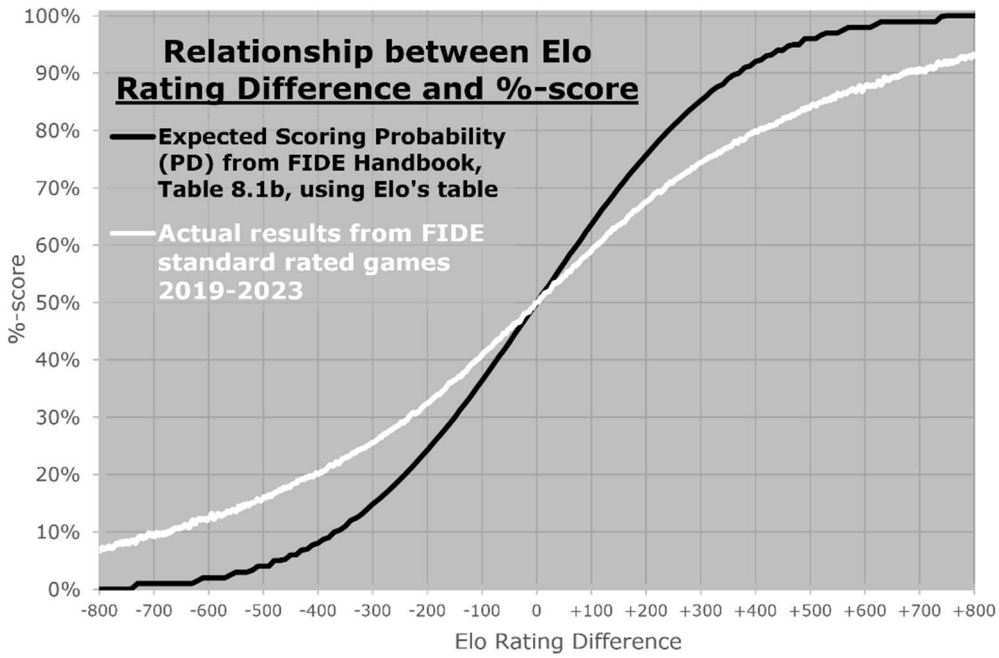
Many chess fans are likely aware that FIDE began contemplating certain changes to its Elo rating system last year, in hopes of effectively tackling the problem of rating deflation. This involved a lot of discussions, proposals, presentations, and meetings. In mid-December, the FIDE Council formally approved the final set of proposed changes, which **will go into effect in March of 2024**. Because most of these changes were my own suggestions, I thought it might be useful for me to share my latest thoughts, and some illustrative analysis, about them.

First, a bit of background and a timeline of how things transpired this past year. I own my own consulting company (Sonas Consulting LLC) and over the past 15 years I have consulted for several different chess organizations, including the FIDE Qualification Commission (the "QC"). Although it had been several years since I last worked with them, the QC once again reached out to me in the spring of 2023, to request another "health check" of the FIDE standard Elo rating system. We assembled a working group within the QC and met several times during the spring and summer to answer questions and discuss my findings, and eventually I submitted a 19-page proposal (https://bit.ly/SonasProposalJuly2023) to senior FIDE management in July of 2023, outlining my recommendations. This proposal was published by FIDE on their website, launching a ten-week-period for public discussion from late July to the end of September, in which members of the chess community were invited (https://bit.ly/SonasProposalNewsJuly) to contribute their thoughts via emails to the QC.

The QC shared all the feedback with me, and after I reviewed all the submissions, performed additional follow-up analysis, and presented my findings in early October to the QC again, they held a vote in which the latest proposals were generally supported, with some minor revisions. I submitted a 47-page supplemental report (https://bit.ly/SonasSupplemental) to FIDE in late October, describing my latest analysis and my final proposal. The QC held an open meeting on December 12th where I gave one last presentation, and the changes were again generally supported by meeting attendees. The FIDE Management Board and FIDE Council also recently approved the changes, and so I believe that means the changes will go into effect in March of 2024.

The video of that QC Open Meeting (with my presentation) is available online (https://youtu.be/nquLRq8C21M), but I also thought it would be useful for me to turn my last presentation into a written article as well, and that's what you are currently reading. I would like to start by summarizing the major problems in the rating system as I see them, and then I will talk about the proposed changes. We begin, as I often do, with the line graph illustrating the fundamental basis of the Elo system.

The FIDE rating handbook provides a table that converts from rating point difference into an expected scoring probability. We can represent this graphically via the black curve that you can see in the following image:



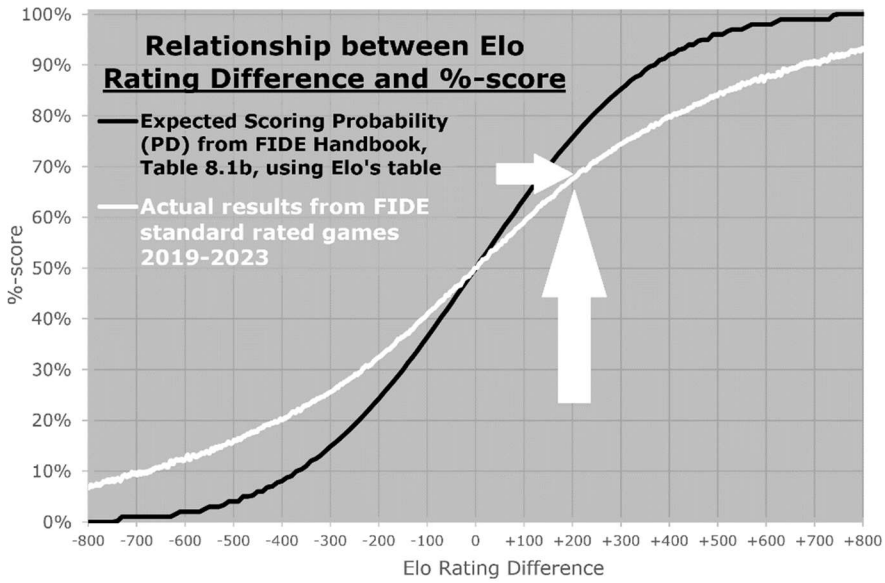
8.1b Table of conversion of difference in rating, D, into scoring probability PD, for the higher, H, and the lower, L, rated player respectively.

D	PD		D	PD	
Rtg Dif	H	L	Rtg Dif	H	L
0-3	.50	.50	198-206	.76	.24
4-10	.51	.49	207-215	.77	.23
11-17	.52	.48	216-225	.78	.22
18-25	.53	.47	226-235	.79	.21
26-32	.54	.46	236-245	.80	.20
33-39	.55	.45	246-256	.81	.19
40-46	.56	.44	257-267	.82	.18
47-53	.57	.43	268-278	.83	.17
54-61	.58	.42	279-290	.84	.16
62-68	.59	.41	291-302	.85	.15
69-76	.60	.40	303-315	.86	.14
77-83	.61	.39	316-328	.87	.13
84-91	.62	.38	329-344	.88	.12
92-98	.63	.37	345-357	.89	.11
99-106	.64	.36	358-374	.90	.10
107-113	.65	.35	375-391	.91	.09
114-121	.66	.34	392-411	.92	.08
122-129	.67	.33	412-432	.93	.07
130-137	.68	.32	433-456	.94	.06
138-145	.69	.31	457-484	.95	.05
146-153	.70	.30	485-517	.96	.04
154-162	.71	.29	518-559	.97	.03
163-170	.72	.28	560-619	.98	.02
171-179	.73	.27	620-735	.99	.01
180-188	.74	.26	> 735	1.0	.00
189-197	.75	.25			

For example, we can look at the table and see that if you have a +200-point rating advantage, you should score 76%, and so at a rating difference of +200 (x-axis), the black curve tells you that indeed you should score 76% (y-axis).

We can also look at what happens in real rated games, in this case what everyone's average percentage scores are at various rating differences, across all of the standard rated games in the past five years. That's what is shown by the white curve. When we do this, we see quite clearly that rating favorites are falling far short of their Elo expectation.

For example, we can look again at +200 rating point advantage, and here we see that the rating favorites are only scoring 68% in real life, rather than 76%:

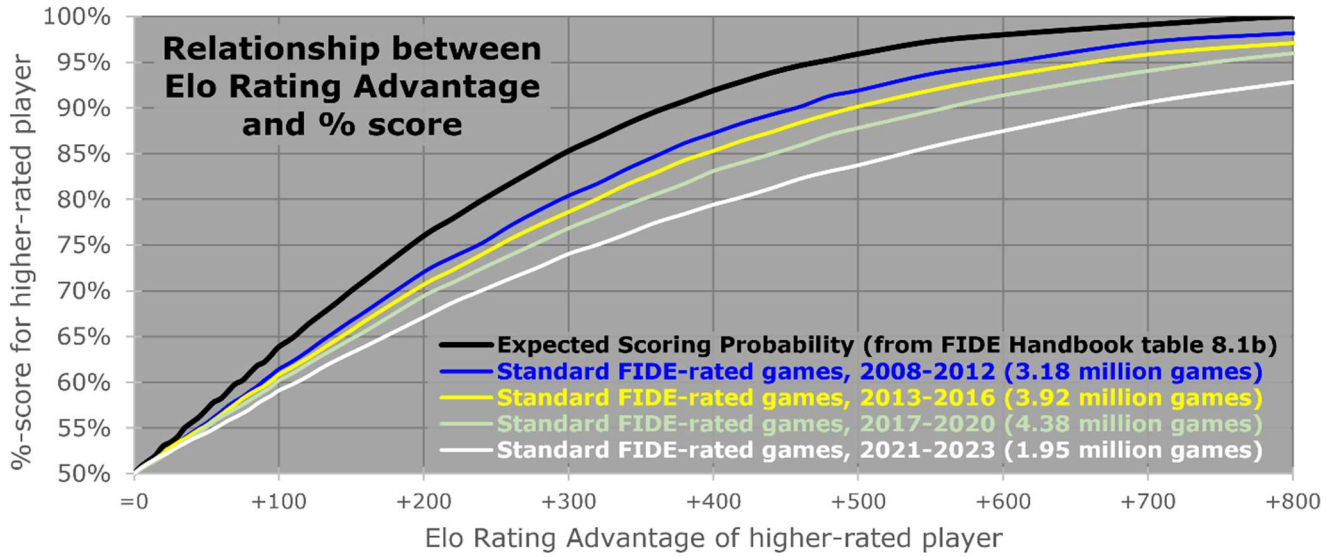


Thus if you have K=20 and you play ten games with a 200-point rating advantage, you can fully expect to lose 15 or 20 Elo points from your rating thanks to those rated games.

Over the years I have drawn this picture several times in QC meetings, or in articles I wrote for ChessBase, to help illustrate this effect. But I have struggled with how best to label the actual effect.

At first I described it as *"rating favorites are underperforming"*, as though it were somehow the fault of the stronger players for not playing well enough. Then I started wondering if the Elo tables were at fault, but when I looked into changing those and did some analysis, it turns out that wouldn't help either. It was only this year, on this project with the QC, that I finally realized the best way to describe what's going on. In fact, the problem is this: *"FIDE Elo ratings are too spread out."* The ratings are exaggerating the differences in strength among players. Indeed, there isn't nearly as large a range of playing strength as the rating list would suggest.

This is not a recent occurrence. COVID-19 made it worse, but COVID-19 didn't cause it. Actually, this behavior is something we noticed a decade or more ago, when I was first brought in by the QC to do health checks on the rating system. However, it has gotten steadily worse over the years, as you can see here from the following historical look. Note that we are zooming into the upper-right quadrant of the previous view, since these curves are symmetrical across the middle, and so you don't get any additional information by seeing the whole thing. So we will just look at the part where the rating advantage ranges from zero to +800, and the percentage-score ranges from 50% to 100%, allowing us to see a bit more detail.



Even back 15 years ago, the blue line was already shallower than the black line, showing that rating favorites were not able to keep up with their expected scores, and thus the ratings were too spread out even then. But it has gotten far worse in recent years. You can see from the yellow (2013-2016), the green (2017-2020), and the white line (2021-2023) that rating differences are becoming less and less meaningful as the years go by.

Admittedly, these curves are describing averages across millions of games, and also combining together results from masters, average players, and junior players. Each little pixel in that last picture might well correspond to hundreds or even thousands of games. Instead, let's try and look at a more manageable example, to give you a better understanding of what's going on, and what it really looks like, practically speaking, for ratings to be "too spread out".

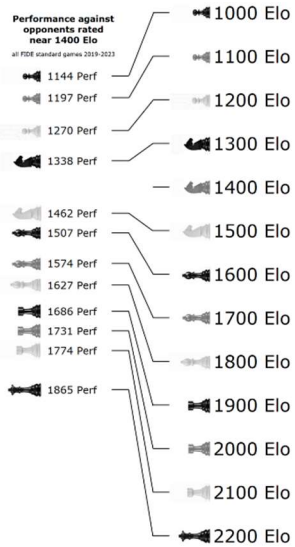
We will focus on the players having ratings near a multiple of 100, so we will look at players rated near 1000 (within +/- 10 Elo points), and near 1100 (within +/- 10 Elo points), and so on. For each of those groups, let's see what happens when they face an opponent rated near 1400. This data spans the past five years of rated games at standard time controls (not rapid or blitz):

Elo Matchup	Expected Score	Actual Score	Rating Perf	Perf + or -
1000 Elo (vs. 1400 Elo)	8%	19%	1144	+144
1100 Elo (vs. 1400 Elo)	15%	24%	1197	+97
1200 Elo (vs. 1400 Elo)	24%	32%	1270	+70
1300 Elo (vs. 1400 Elo)	36%	41%	1338	+38
1500 Elo (vs. 1400 Elo)	64%	59%	1462	-38
1600 Elo (vs. 1400 Elo)	76%	65%	1507	-93
1700 Elo (vs. 1400 Elo)	85%	73%	1574	-126
1800 Elo (vs. 1400 Elo)	92%	79%	1627	-173
1900 Elo (vs. 1400 Elo)	96%	84%	1686	-214
2000 Elo (vs. 1400 Elo)	98%	88%	1731	-269
2100 Elo (vs. 1400 Elo)	99%	90%	1774	-326
2200 Elo (vs. 1400 Elo)	100%	95%	1865	-335

So first of all, against 1400-level opponents, the results for players rated below 1400 are all highlighted here in blue, and they are all overperforming their rating expectation. For example, notice the top row of blue data. 1000-rated players are expected to score 8% against 1400-level opponents, and instead they are scoring 19%, which is a performance rating of 1144. 1100-rated players are also over-performing with a performance rating of 1197, and so on.

Conversely, when we look at how the higher-rated players are performing against 1400-level opposition, they are all underperforming. All of these higher-rated groups are highlighted in red. Every group scores worse than their Elo expectation.

In that listing of player groups over on the far left, you can see that players rated near 1000 Elo are shown as black pawns, players rated near 1100 Elo are shown as gray pawns, etc., and each group is shown equally spaced on the vertical range from 1000 Elo to 2200 Elo. That's what their published Elo ratings were. But then for each of these groups, we can look at their performance rating when facing 1400-level opposition. And for each one, we will now draw them at the appropriate place on that same scale.

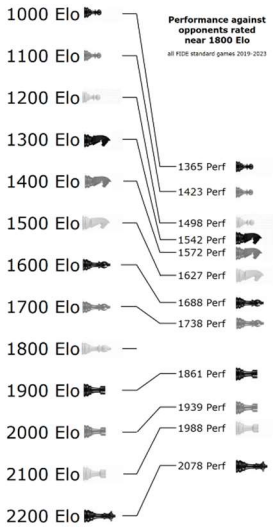


Elo Matchup	Expected Score	Actual Score	Rating Perf	Perf + or -
1000 Elo (vs. 1400 Elo)	8%	19%	1144	+144
1100 Elo (vs. 1400 Elo)	15%	24%	1197	+97
1200 Elo (vs. 1400 Elo)	24%	32%	1270	+70
1300 Elo (vs. 1400 Elo)	36%	41%	1338	+38
1500 Elo (vs. 1400 Elo)	64%	59%	1462	-38
1600 Elo (vs. 1400 Elo)	76%	65%	1507	-93
1700 Elo (vs. 1400 Elo)	85%	73%	1574	-126
1800 Elo (vs. 1400 Elo)	92%	79%	1627	-173
1900 Elo (vs. 1400 Elo)	96%	84%	1686	-214
2000 Elo (vs. 1400 Elo)	98%	88%	1731	-269
2100 Elo (vs. 1400 Elo)	99%	90%	1774	-326
2200 Elo (vs. 1400 Elo)	100%	95%	1865	-335

So even though all these players span a range of 1200 Elo points in their ratings, that is not what we see from their performance; they are clustered together much more closely than that, and in this case barely spanning a range of 700 Elo points in their actual performance, from the black pawns having Elo ratings near 1000 and performing at 1144 Elo, all the way up to the black queens having Elo ratings near 2200 and performing at 1865 Elo against such opponents.

And there is nothing special about 1400-level opponents. Let's try a different level; we will look at 1800-level opponents instead. Once again, we will highlight the performances of the rating underdogs in blue, and the rating favorites in red, and we will draw the performances of each group in their proper place on the overall scale from 1000 Elo up to 2200 Elo, when facing 1800-level opponents:

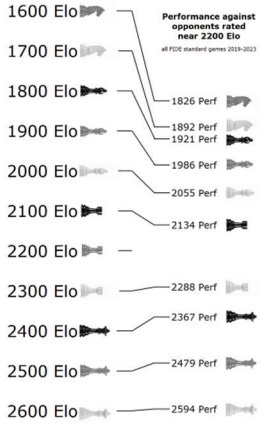
Elo Matchup	Expected Score	Actual Score	Rating Perf	Perf + or -
1000 Elo (vs. 1800 Elo)	0%	6%	1365	+365
1100 Elo (vs. 1800 Elo)	1%	9%	1423	+323
1200 Elo (vs. 1800 Elo)	2%	15%	1498	+298
1300 Elo (vs. 1800 Elo)	4%	18%	1542	+242
1400 Elo (vs. 1800 Elo)	8%	21%	1572	+172
1500 Elo (vs. 1800 Elo)	15%	27%	1627	+127
1600 Elo (vs. 1800 Elo)	24%	35%	1688	+88
1700 Elo (vs. 1800 Elo)	36%	41%	1738	+38
1900 Elo (vs. 1800 Elo)	64%	59%	1861	-39
2000 Elo (vs. 1800 Elo)	76%	69%	1939	-61
2100 Elo (vs. 1800 Elo)	85%	74%	1988	-112
2200 Elo (vs. 1800 Elo)	92%	83%	2078	-122
2300 Elo (vs. 1800 Elo)	96%	87%	2127	-173
2400 Elo (vs. 1800 Elo)	98%	90%	2174	-226



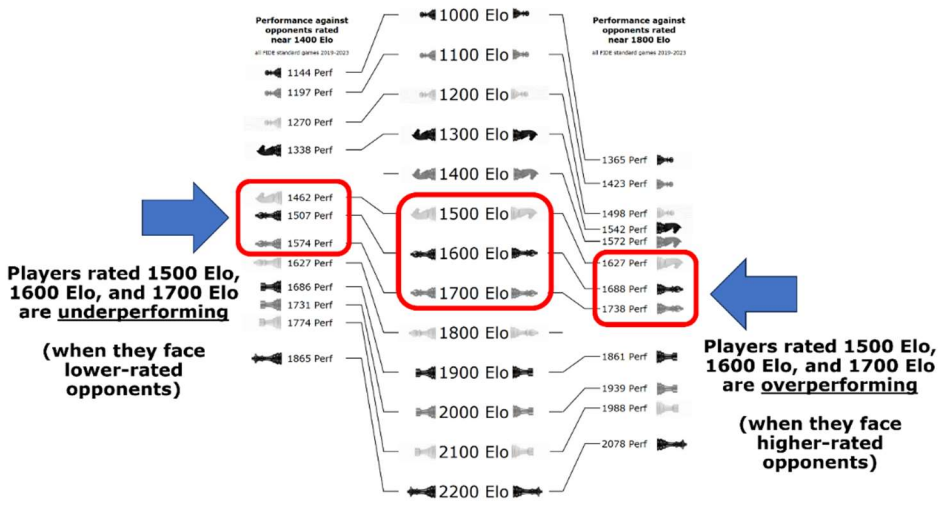
It's the same thing again. All of the players rated below 1800 are highlighted in blue, and you can see they overperformed their ratings when facing 1800-level opposition. And conversely, the players rated higher than 1800 are highlighted in red, and performed worse than their ratings would suggest. So once again, when we line up everyone based on their rating performance (see the far right), they are clustered more closely together, barely spanning 700 Elo points of performance, despite spanning a full 1200 Elo points on the rating list.

It turns out that it doesn't matter much what level of opposition you choose. It could be 1400, 1800, even 2200 has the same type of behavior, although games played among masters don't deviate quite as much from expectation as what we see lower down in the rating list. For the sake of completeness, here is what it looks like against 2200-level opponents, restricting ourselves only to the matchups with a reasonable amount of games played:

Elo Matchup	Expected Score	Actual Score	Rating Perf	Perf + or -
1600 Elo (vs. 2200 Elo)	2%	10%	1826	+226
1700 Elo (vs. 2200 Elo)	4%	14%	1892	+192
1800 Elo (vs. 2200 Elo)	8%	17%	1921	+121
1900 Elo (vs. 2200 Elo)	15%	23%	1986	+86
2000 Elo (vs. 2200 Elo)	24%	31%	2055	+55
2100 Elo (vs. 2200 Elo)	36%	41%	2134	+34
2300 Elo (vs. 2200 Elo)	64%	62%	2288	-12
2400 Elo (vs. 2200 Elo)	76%	72%	2367	-33
2500 Elo (vs. 2200 Elo)	85%	84%	2479	-21
2600 Elo (vs. 2200 Elo)	92%	92%	2594	-6

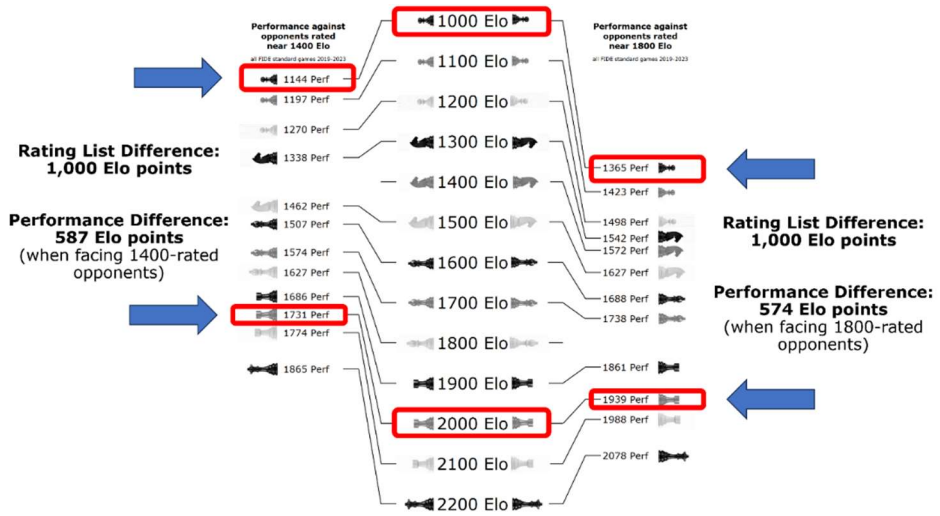


Let's go back to the diagrams representing performances against 1400-level and 1800-level opponents. We shall place them side-by-side so we can talk about a couple more things:



I want to emphasize that I'm not saying entire groups are overrated or underrated. For example, let's look at the three groups shown in the center, outlined in red. Namely, the players with ratings near 1500, or near 1600, or near 1700. They are under-performing when they face weaker opponents, but on the other hand they are over-performing when they face stronger opponents. I am trying to emphasize here that player ratings have gotten too separated, rather than targeting individual groups and calling them overrated or underrated.

And one last, very important thing:



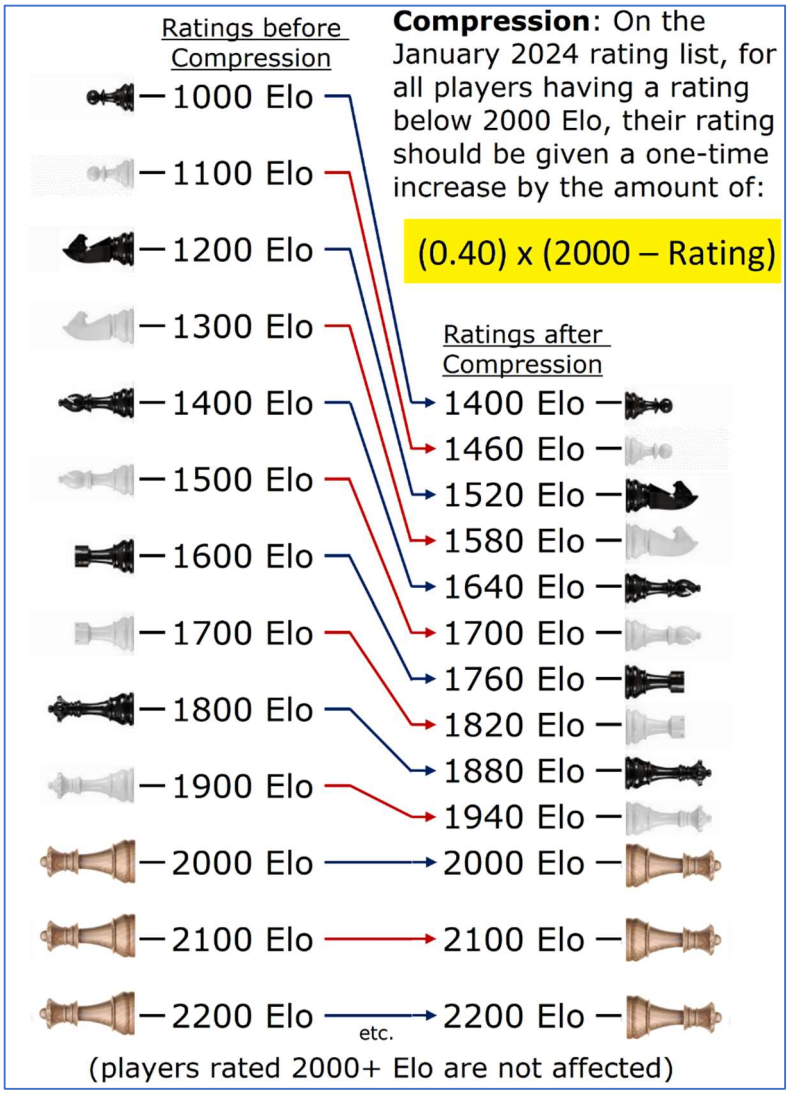
Let's look at the players who are shown as black pawns, and outlined in red near the top. These are the ones having a FIDE rating near 1000. When facing 1400-level opponents (on the left), they are performing around 1144. And if we look down below, at the players who are shown as gray rooks and outlined in red near the bottom, they have FIDE ratings near 2000, and they are performing around 1731 against 1400-level opponents (also on the left). So despite the ratings claiming that the difference in playing strength between the black pawn players and the gray rook players is 1,000 Elo points, the actual games from the past five years make it look more like they are barely 600 Elo points apart – if that! - in terms of their evident playing strength.

And next, focus on the right side. This same basic thing is true if we look at games against 1800-level opponents. For the black pawns and gray rooks, their ratings are different by 1000 Elo points, but their differences in performance aren't quite 600 Elo points.

So this is one way of trying to measure how much the ratings are too spread out. And I've tried eight or ten different ways of looking at this - note that I go into much more detail about other approaches, in that 47-page supplemental report - and it always says roughly the same thing. Wherever you look, among these players rated from 1000 Elo up to 2000 Elo, the difference in performance is only about 60% of the difference in published ratings. This is what led me ultimately to suggest a compression, where we take all the players rated from 1000 to 2000, and squish them closer together so that they only span 600 Elo points instead of spanning 1,000 Elo points. And because the stretchedness of the ratings is quite consistent across this span from 1000 Elo to 2000 Elo, the compression can be done uniformly, based upon a simple linear formula.

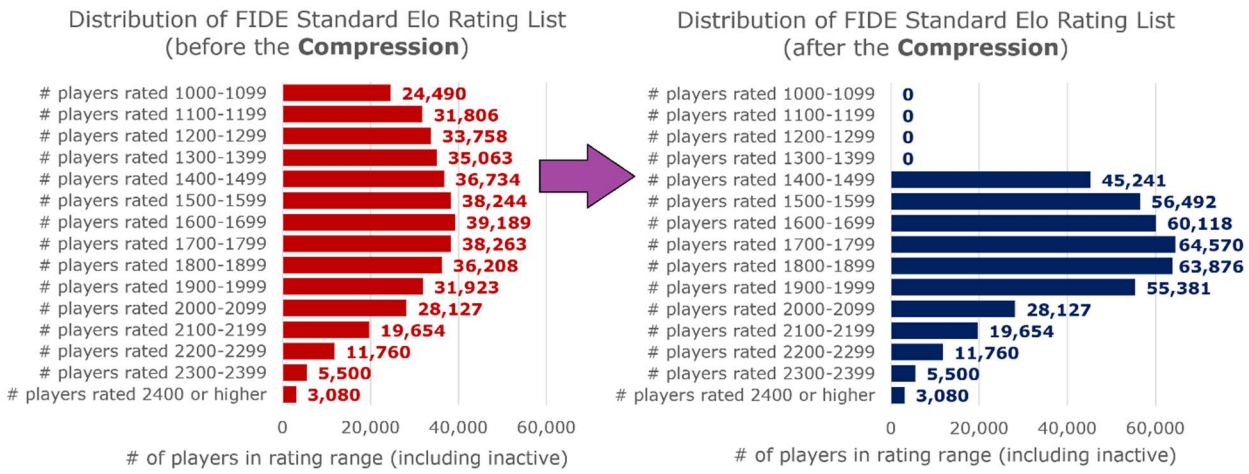
I described my proposed changes as **"Compression and Calculation Improvements"**. Essentially, we have a strong deflationary effect in the rating system that is constantly operating, a downward force being introduced at the lower end of the rating pool that is tugging everyone's ratings gradually lower and lower. This is ultimately causing the ratings to get more stretched apart, to the point that today's ratings are greatly exaggerating the differences in playing strength among players. Because the ratings are expanding downward like this, I have used the term "deflation" to describe the effect, but "downward expansion" is probably more accurate.

I have proposed a two-pronged approach to dealing with the situation. The first priority is to adjust the ratings so they are more in line with the evident range of playing strength that we actually see in real-life chess. The second priority is to make sure we understand how this came to be, and try to keep it from happening again. Let's start with the first priority, the Compression.



So as illustrated in the graphic above and to the right, I propose that we take everyone rated 2000 or below, and give them a one-time rating bonus which is based on how far away they are from 2000 Elo. Players down near 1000 Elo will get about 400 points added to their rating, players around 1500 Elo will get about 200 points added to their rating, and players just below 2000 will get almost nothing added to their rating.

Here is an example of what this change might look like, in terms of the whole distribution of players:

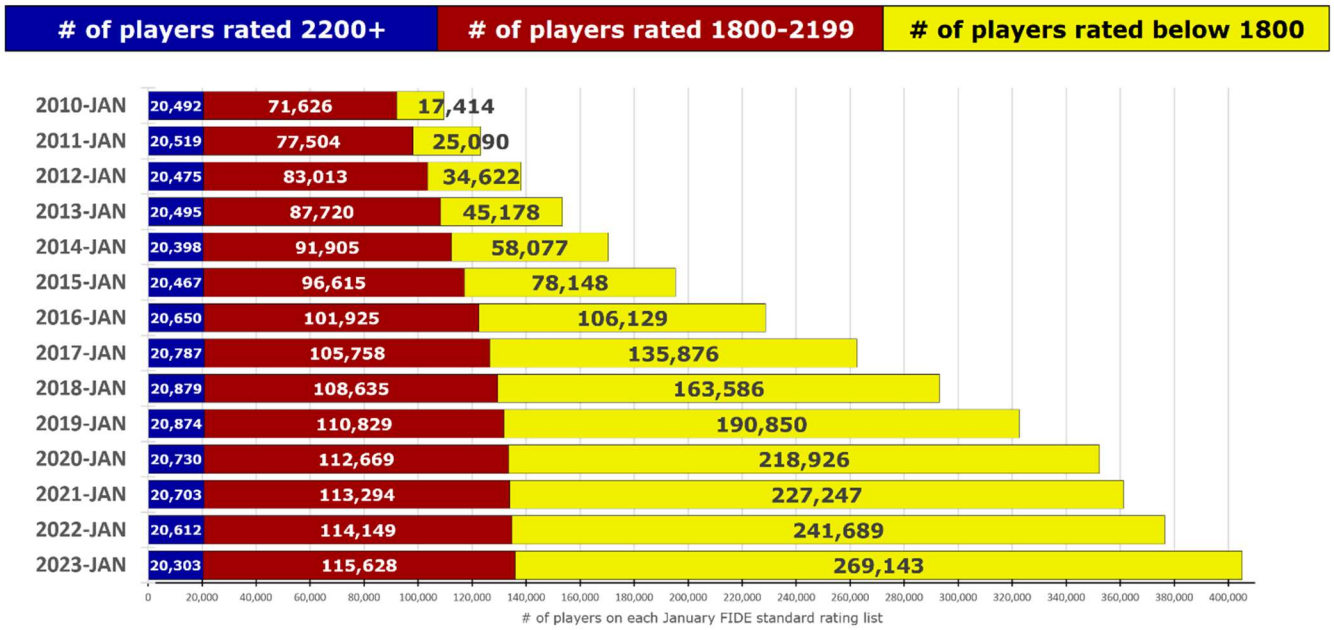


On the red bars on the left, we see how many players are rated in the 1000's, in the 1100's, and so on. And then on the right side, in dark blue, we see what it would look like after the compression.

Because everybody would rise to 1400 Elo or higher, there would be no players left in the 1300's, 1200's, or lower. And the rating ranges of 1400's, 1500's, etc., up to 1900's, would get more crowded. There would be no additional players with Elo ratings of 2000 or higher, but we would expect an increase like that in the next few months after the compression, because players would more easily cross over 2000 Elo now.

So let's move on to what is probably the harder question, namely how we got into this situation, and is there some way we can prevent it happening again? I should also mention that in contemplating possible solutions, I considered it a high priority to retain the simplicity of the existing Elo system as much as possible. Any complexity that we add, any extra rules we need to introduce, could move us further and further away from it being so easy for chessplayers of all mathematical prowess to calculate their own ratings. This was quite a strong constraint to adopt, but it seemed like a worthy one, in the spirit of the Elo system.

So far we have been looking mostly at data about game results, but let's switch over to looking at the composition of the rating list itself. On the following bar charts, I took each January rating list, going all the way back to 2010, and just counted up how many rated players there were in different Elo ranges, and that includes both active and inactive players:



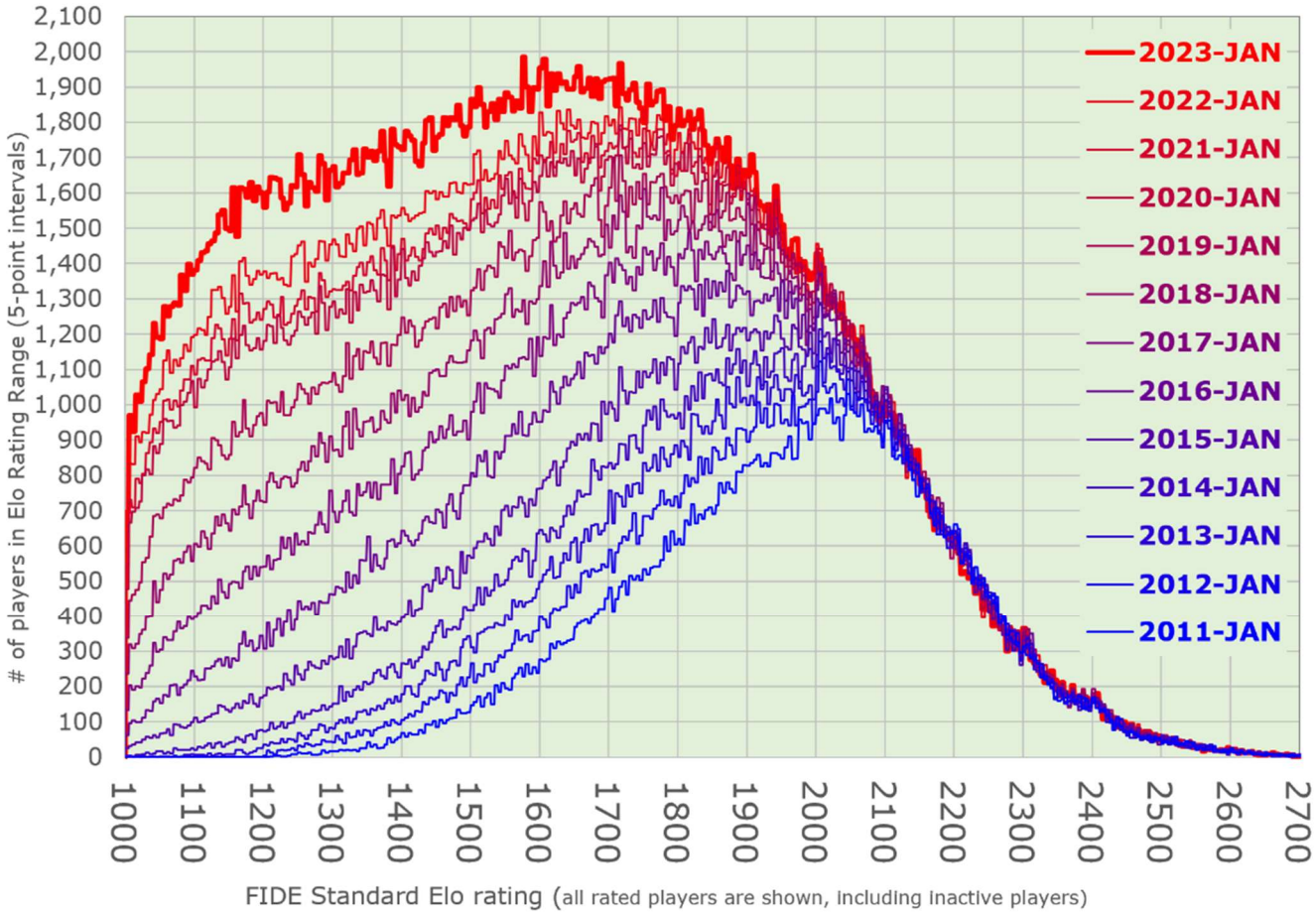
On each January list, the blue tells us how many players were rated 2200 or higher, the red tells us how many were rated from 1800 to 2199, and the yellow tells us how many were below 1800. These are some very wide ranges of Elo ratings, with the narrowest one still being 400 Elo points wide, but what it lacks in detail, it will make up for in clarity; it tells a very clear picture.

Two things kind of leap out at you from this. First of all, we have had fantastic growth in the yellow group, the lowest-rated players. The yellow bars are just getting bigger and bigger each year. And second, the blue bars are not changing at all! Every single year going back to 2010, the January list has had between 20,000 and 21,000 players with master-level ratings, meaning 2200 Elo or higher.

And all of those players who were in the strongest "blue" group in 2010 – as long as they are still alive, they are still on the rating list today, in 2023/2024! And so they would still count toward the numbers for the blue group today, unless they lost so many rating points that they are over in the red group now.

To me, this just doesn't quite add up. This can't be the true strengths of all these players, because we've had this gigantic influx of scholastic players coming into the yellow group, year after year, and somehow almost none of them have risen to master-level strength, even after all these years? And also, that blue group has actually now been shrinking every January for the past five years. As expected from those trends, we only barely stayed above 20,000 on the 2024-JAN list (the actual number was 20,088). So really I think this is exactly what it looks like for deflation to have finally reached the top, when your overall pool of players is still growing so healthily. First the top group stops growing very much, then it starts shrinking. And we are at the shrinking stage right now.

We can also look at the annual January distributions of player ratings in a much more detailed way, by looking at Elo ranges that are only 5 points wide, instead of 400 points wide, and then drawing a histogram of them for each year's January list.



In this picture, the blue line, down near the bottom, tells us what the distribution of FIDE ratings was in January 2011, and then each additional line above that is for one year later, and gradually moves from blue to purple to red, until we reach January 2023's rating distribution, which is the solid red color at the top. So again, if you look at the left half of this picture, we see that in the weaker part of the rating list, there has been impressive growth, especially starting in 2015 when it became easier for players to get an initial rating.

But what's amazing is what's happened over on the right side, where there is zero growth at the master level and higher. In fact, you can even see some spots (especially in the 2200-2300 range) where the red is sometimes peeking out at the bottom, meaning we are currently seeing fewer players in that Elo range than at any point in the past dozen years!

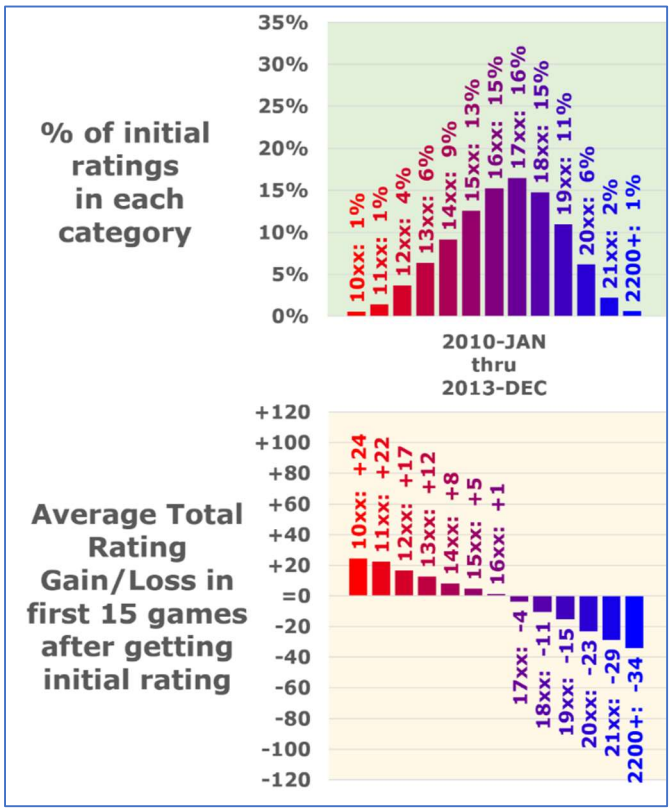
My contention is that there is a strong deflationary effect, made worse by having so many players who are weak, young, improving, and underrated, entering the rating pool and then immediately taking rating points away from established players. Those opponents in turn then become slightly underrated, and they take a few rating points away from their next opponents, and eventually this effect propagates throughout the entire rating pool, pulling everyone's ratings down.

I want to draw a distinction there between "improving" and "underrated". We all know that junior players tend to improve, so it's not a big surprise if they start gaining rating points once they get a rating. However, I believe that these newly rated players, on average, are also

immediately underrated, the very moment they get their first rating, and so they don't even need to start improving; they will gain Elo points just by playing at their current ability level.

In any event, while this effect has gradually propagated throughout the rating pool, it's taken longest to make its way to the super-GM's, since they mostly play each other and so it doesn't impact them as much if some lower-rated masters start getting underrated a bit. But ultimately all the players are part of the same connected network of opponents, and ultimately the deflation has even reached the 2600's now. For several consecutive months, I've downloaded the latest FIDE rating list and found that yes indeed, there's fewer players rated 2600+ on that month's list than in the previous month. I talk a lot more about the deflation at the highest levels, in my 47-page supplemental report that I've mentioned previously.

In my opinion, the major culprit for the deflation is the formula that calculates initial ratings. I have gone through the following pictures with the QC, but I don't think it's been made public previously. It makes it quite easy to understand where the deflation has originated.



First we will start with how things were about a dozen years ago, specifically 2010-JAN through 2013-DEC. In the top half of the picture at the right, you can see several columns, ranging from the 1000's and 1100's (in red) over to the 2100's and 2200's (in blue). Each vertical bar tells you what the relative frequency was, for new ratings in that Elo range across those four years. For example, the tallest bars are for the 1600's, 1700's, and 1800's. 15% of all new ratings were in the 1600's, 16% of all new ratings were in the 1700's, and 15% of all new ratings were in the 1800's. Adding those three together, we see that 46% of all new ratings from 2010-2013 were somewhere between 1600 Elo and 1899 Elo.

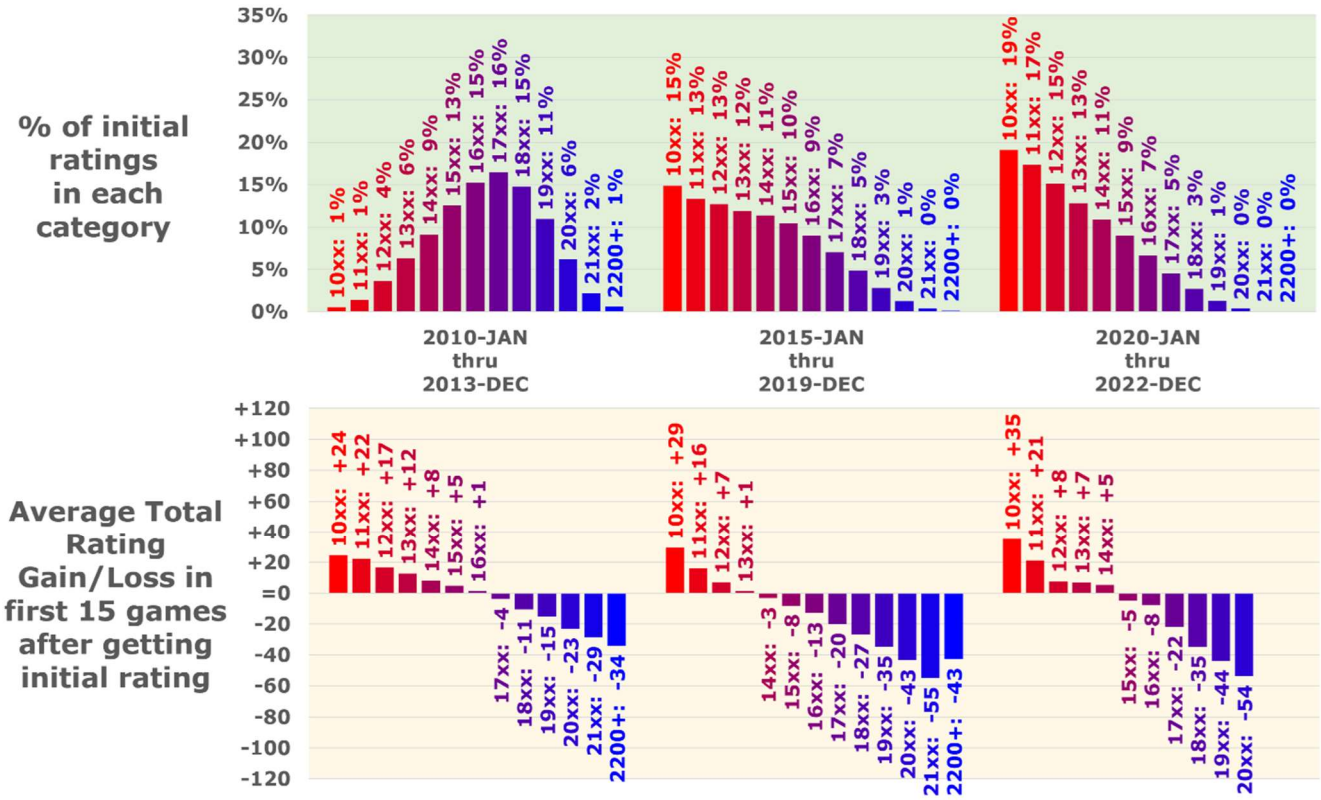
For each bar, you can look down and see another bar directly below it. Either that bar reaches upward, indicating a positive number, or it reaches downward, indicating a negative number. What you're looking at there is the average rating gain (or loss) in the first 15 games that a player has after they get their first rating. For example, you can see for 11xx that the number it shows is +22. This means on average, players who got an initial rating from 1100-1199 were able to gain 22 Elo points across their first 15 games as a rated player (during 2010-2013). That suggests that those ratings were a bit low, since an average rating gain like that is quite high. It means those players were immediately scoring about 0.05 higher per game than their Elo expectation, directly after getting their first rating.

However, look over to the right, at the blue bars. Amazingly, we see that players who got initial ratings in the 2000's, 2100's, etc., were actually losing rating points, on average, in their first 15 games after getting that initial rating. This is quite striking, because we would expect that many such players were young and still improving, and so it wouldn't make sense that they would be immediately losing rating points on average, unless their initial rating had been too high. But the trend is quite clear there - the very low initial ratings were probably (on average) too low, and the very high initial ratings were probably (on average) too high.

In fact, we were actually aware of all of this in our QC meetings, a dozen or more years ago, and struggled with what to do about it. At the time, I suggested that we add a couple of hypothetical draws into the game results that were used for calculating players' initial ratings. If everyone had two draws against 1600-level opponents, that would pull everyone's initial ratings more toward the center, more in line with their anticipated performance in the near future. But there were concerns about this being kind of arbitrary, and where should this magic 1600 number come from, and eventually... we decided not to do it. In retrospect, and also based on some of my recent simulations, I actually think that we would be in a much better place today with the ratings, if we had gone ahead and made that change back then. The theoretical basis for it could be that we have a prior guess at players' strength, before knowing much about their initial results, and so that prior estimate gets represented by those two results of an average player's performance, and that gets merged with the evidence of a small number of their actual game results, to determine their initial rating.

So back then, there was a sort of equilibrium in place. A bunch of players getting new ratings that were too low, and then taking points away from the rating pool. And a bunch of players with new ratings that were too high, who were then losing points to the rating pool. Essentially, they mostly canceled each other out, and had no great impact on the established rating pool.

Enough about those times gone by. Let's move more into the present, and add a couple more eras, one being 2015-JAN through 2019-DEC, and the other being 2020-JAN through 2022-DEC:

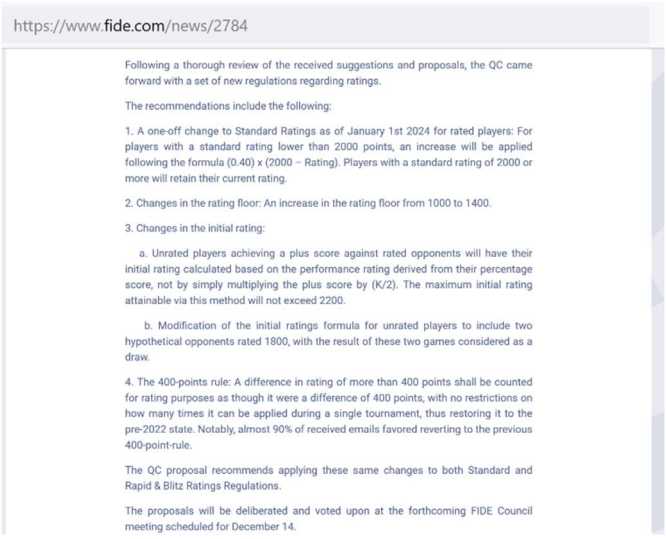


And here we see the problem. There is still the same trend in the bottom half of the picture, showing that the initial ratings are too spread out (players with low initial ratings quickly gaining rating points, and players with high initial ratings quickly losing rating points). But if you look up at the top of the picture, the balance is no longer there. There's far more players getting low initial ratings than there are players getting high initial ratings. And in fact, the most frequent initial ratings are in the 1000-1099 and 1100-1199 groups, which are the ones showing by far the biggest initial improvement, once those players start facing the established rating pool.

The net result is an ongoing system of new ratings that are inflicting a strong and constant deflationary effect upon the established rating pool. It's almost like a "deflation engine". Each month there are more players coming in, and then (on average) immediately taking more points away from the pool. So we have ongoing deflation. Or perhaps it's better to call it downward expansion of ratings, but either way, it never stops.

There is a school of thought that the Elo system is self-correcting, and maybe we just need to be patient and let it do its thing, but I feel very strongly that we cannot just be patient and hope it gets better with the passage of time. We are continually getting new players entering the system, and with progressively higher playing strength corresponding to any given initial rating level. Thus the system is constantly being disrupted and ratings pulled downward. I am quite confident there will never be time for the Elo system to right itself. It will just keep getting worse and worse, unless we break this pattern and make some major changes to counteract the existing deflation and the ongoing deflationary effect.

So that's what led directly to my suggestion this year, that we adjust the formula used for the calculation of initial ratings, so that it incorporates two hypothetical draws against 1800-level opposition. By pulling everyone's initial rating closer to the center, we will hopefully be reducing the impact of newly rated players upon the existing rating pool. I arrived at 1800 by doing various simulations that considered different numbers - as though these changes had been deployed years ago, and then looking at what ratings would be like today - and 1800 seemed to behave the best. It may seem like an overly high choice, but remember that it's more like 1667 by today's standards, and also there is a beneficial systemic effect to having a slightly high number like this, as it is another factor that pushes back slightly against deflation.


https://www.fide.com/news/2784

Following a thorough review of the received suggestions and proposals, the QC came forward with a set of new regulations regarding ratings.

The recommendations include the following:

1. A one-off change to Standard Ratings as of January 1st 2024 for rated players: For players with a standard rating lower than 2000 points, an increase will be applied following the formula (0.40) x (2000 – Rating). Players with a standard rating of 2000 or more will retain their current rating.

2. Changes in the rating floor: An increase in the rating floor from 1000 to 1400.

3. Changes in the initial rating:

a. Unrated players achieving a plus score against rated opponents will have their initial rating calculated based on the performance rating derived from their percentage score, not by simply multiplying the plus score by (K/2). The maximum initial rating attainable via this method will not exceed 2200.

b. Modification of the initial ratings formula for unrated players to include two hypothetical opponents rated 1800, with the result of these two games considered as a draw.

4. The 400-points rule: A difference in rating of more than 400 points shall be counted for rating purposes as though it were a difference of 400 points, with no restrictions on how many times it can be applied during a single tournament, thus restoring it to the pre-2022 state. Notably, almost 90% of received emails favored reverting to the previous 400-point-rule.

The QC proposal recommends applying these same changes to both Standard and Rapid & Blitz Ratings Regulations.

The proposals will be deliberated and voted upon at the forthcoming FIDE Council meeting scheduled for December 14.

I also experimented with different simulations in which, after the Compression that raises all existing player ratings to a minimumn of 1400 Elo, we try different approaches for a new minimum rating; either we leave the minimum as 1000 Elo, or we raise the minimum to 1200 Elo, or we raise the minimum to 1400 Elo. What I found was that the deflation returns quite rapidly if we keep those bottom ranges open, and so it seemed most prudent to seal off the minimum rating at 1400 Elo.

Some people expressed an understandable concern that we would be blocking many deserving new players from getting ratings, if we chose to raise the minimum rating to 1400 Elo. However, in my opinion this concern is unfounded.

Remember that all existing rated players are getting their ratings increased to at least 1400 Elo by the compression. In 2023, if there was a new player who performed slightly better than the lowest-rated players in the current rating list, that new player would just sneak onto the bottom of the rating list, with a new rating slightly above the 1000 minimum. Similarly, after the compression and calculation improvements, a new player who performs slightly better than the lowest-rated players in the rating list, would just sneak onto the bottom of the rating list with a new rating slightly above the 1400 minimum. And it's the same players who are at the bottom of the rating list, so it's no more exclusive or inclusive for new rated players than it previously was. Today's 1000-rated player is tomorrow's 1400-rated player. Truthfully, the new system will be more inclusive, due to the two hypothetical draws pulling some initial ratings up above 1400.

There is also a proposed change for initial ratings so that if the unrated player gets a plus score while unrated, they receive their actual performance rating (counting the two extra draws) instead of the previous formula, which only gave you a rating of your opponents' average rating plus (K/2) times your plus score. This ought to provide a mild counter to the deflation as well, and a maximum initial rating of 2200 Elo ought to hold back overly-high initial ratings.

During my simulations, all the proposed changes taken together did seem to do a pretty good job of holding back the deflation. It is quite possible, however, that we will eventually need to take additional steps. The biggest vulnerability will probably be that even though new junior players are getting K=40 and so their ratings will go up a lot as they overperform their expectation, most of their opponents are also K=40 juniors who will lose a corresponding amount of rating points from such games. So even though both players are improving, no rating points are introduced into the rating pool. It may become necessary to take measures such as the Chess Scotland (CS) system uses, where if two equally-rated juniors face each other and draw the game, both players gain rating points from the outcome. This would certainly add some complexity to the rating calculation, so it would be ideal if we didn't have to do this.

The last of my suggestions, and perhaps the most controversial, is to revert the 400-point-rule back to how it was a couple of years ago. In the current incarnation of the rule, a player can only benefit once per tournament from having their expected score reduced down to 92%, even if they have multiple games with a rating advantage of 412+ Elo points. If we revert the rule to its pre-2022 state, there would once again be no "once per tournament" restriction to this rule.

I know that the current version of the rule was introduced essentially as an anti-cheating measure, to prevent strong titled players from "harvesting" rating points by playing in very weak events where they seem guaranteed to score 100% and gain 0.8 rating points per game played. However, the counterpoint to this is that the current version of the rule seems to impose an undue burden onto rating favorites who are not even able to manage a 92% score in such games, let alone something higher. And that latter group is far larger than the group of IM/GM's whose potential rating abuse we would be trying to limit via the current rule.

To gain some further insight into the situation, let's look at some numbers. Despite the rule's name, the only games that are pertinent to the 400-point-rule are the games with an Elo difference of 412 points or more, since that's when the expected score reaches 93% or higher in the Elo table; see that portion of the table, below:

Table of conversion of difference in rating, D, into scoring probability PD, for the higher, H, and the lower, L, rated player respectively.

D	PD	
Rtg Dif	H	L
345-357	.89	.11
358-374	.90	.10
375-391	.91	.09
392-411	.92	.08
412-432	.93	.07
433-456	.94	.06

I invented a term **"Extreme Elo Difference"** to refer to just those games. And what I did was to look back at the results of Extreme Elo Difference games during various eras since the 400-point-rule was added, to see if there is evidence of stronger players unfairly benefiting or suffering (on average) from the 400-point-rule, during those years.

For each era, we can go back through all of its Extreme-Elo-Difference games and see what the expected score would have been, if various versions of the 400-point-rule had been in place. For example, under the traditional 400-point-rule, the expected score for all these games would be 92%. Or if the "max-1-upgrade" rule had been in effect, the expected score would usually be

around 93% or 94%. And finally, there would be the option of not having a 400-point-rule at all, in which case we can see what the expected score would have been, just by looking up the difference in the Elo table. That will be even higher than the expected score for the "max-1-upgrade" rule.

As an example, let's look back at the first couple of years after it went from a 350-point-rule to a 400-point-rule, namely 2010 and 2011. There were about 72,000 rated games that count as Extreme Elo Difference games during those two years, and about 17,000 of them, roughly a quarter of these games, involved the higher-rated player being rated 2400-2799 Elo, so about IM or GM level. The overall stats for those 17,000 games are shown in the graphic to the right.

	Years 2010 and 2011
When higher-rated player is rated 2400-2799 Elo	17,090 games (24% of total) actual: 95.4% score traditional rule: 92% expect 1-upgrade rule: 93.3% expect **no rule: 96.1% expect**
LEGEND:	**YELLOW** indicates a matchup where completely removing the 400-point-rule does best at matching the actual performance results

On average, the titled player scored more than 95% in those games, although the official expected score was only 92%. Even the "max-1-upgrade" rule wouldn't have been good enough, since the expected score would have been just 93.3% under that rule. So the best approach for handling the performance of IM and GM players back then, would've been having no 400-point-rule at all (with an expected score of 96.1%), since the IM's/GM's were able to exploit the 400-point-rule pretty significantly back then. And that's why this cell is color-coded yellow, to show that the best option (for those years and that matchup) would have been eliminating the rule completely.

Next, let's move six years into the future, looking at 2016 and 2017:

	Years 2016 and 2017
When higher-rated player is rated 2400-2799 Elo	21,549 games (8% of total) actual: 94.2% score traditional rule: 92% expect **1-upgrade rule: 93.8% expect** no rule: 96.5% expect
LEGEND:	**RED** indicates a matchup where the current 400-point rule (maximum of 1 upgrade to 92%) from 2022 does best at matching the actual performance results

By now, probably to some degree thanks to deflation, these strongest players (the ones rated 2400-2799 Elo) were barely able to score 94% in these games. In fact, the best formula for expected score, six years ago, would actually have been the rule in its current state, where a player can only benefit from (at most) one upgrade each tournament. As you can see, its expected score of 93.8% is the closest to the actual score of 94.2%, out of the different 400-point-rule variations under consideration. That's why this cell is color-coded red.

And finally we can move to the present day, with games from 2022 and 2023:

	Years 2022 and 2023
When higher-rated player is rated 2400-2799 Elo	10,259 games (5% of total) actual: 92.1% score **traditional rule: 92% expect** 1-upgrade rule: 94.2% expect no rule: 96.8% expect
LEGEND:	**BLUE** indicates a matchup where the traditional 400-point-rule (unlimited upgrades to 92%) does best at matching the actual performance results

Here we find that once again, the strongest players are not dominating the lower-rated opponents as easily as before, only managing a score of 92.1% against them in these games. So in recent years, the traditional 92% rule would actually have worked best of all, with an expected score much closer than is the case for the "max-1-upgrade" rule or having no rule at all. And that's why this cell is color-coded blue.

So depending on which era you look at, a different variation of the 400-point rule seems best, at least when the higher-rated player is IM/GM-level. However, the rating list spans such a large range of ratings that the 400-point rule applies to a lot of lower-rated rating favorites as well, even untitled rating favorites. We can use the same color-coding scheme, but expand out to include the matchups involving the lower-rated rating favorites too. When we color-code all of those scenarios, the story becomes much clearer:

Extreme-Elo-Difference games (those having Elo difference of 412+)

	Years 2010 and 2011	Years 2016 and 2017	Years 2022 and 2023
When higher-rated player is rated below 2000 Elo	11,849 games (16% of total) actual: 88.9% score **traditional rule: 92% expect** 1-upgrade rule: 92.3% expect no rule: 95% expect	138,849 games (53% of total) actual: 88.0% score **traditional rule: 92% expect** 1-upgrade rule: 93% expect no rule: 95.9% expect	114,325 games (61% of total) actual: 83.2% score **traditional rule: 92% expect** 1-upgrade rule: 93.3% expect no rule: 96.1% expect
When higher-rated player is rated 2000-2399 Elo	43,691 games (60% of total) actual: 92% score **traditional rule: 92% expect** 1-upgrade rule: 92.9% expect no rule: 95.7% expect	99,510 games (38% of total) actual: 91% score **traditional rule: 92% expect** 1-upgrade rule: 93.6% expect no rule: 96.4% expect	63,172 games (34% of total) actual: 87.1% score **traditional rule: 92% expect** 1-upgrade rule: 94.2% expect no rule: 96.9% expect
When higher-rated player is rated 2400-2799 Elo	17,090 games (24% of total) actual: 95.4% score traditional rule: 92% expect 1-upgrade rule: 93.3% expect **no rule: 96.1% expect**	21,549 games (8% of total) actual: 94.2% score traditional rule: 92% expect **1-upgrade rule: 93.8% expect** no rule: 96.5% expect	10,259 games (5% of total) actual: 92.1% score **traditional rule: 92% expect** 1-upgrade rule: 94.2% expect no rule: 96.8% expect
Total	72,630 games total actual: 92.3% score **traditional rule: 92% expect** 1-upgrade rule: 92.9% expect no rule: 95.7% expect	259,908 games total actual: 89.7% score **traditional rule: 92% expect** 1-upgrade rule: 93.3% expect no rule: 96.1% expect	187,756 games total actual: 85% score **traditional rule: 92% expect** 1-upgrade rule: 93.6% expect no rule: 96.4% expect

LEGEND:	**BLUE**	indicates a matchup where the traditional 400-point-rule (unlimited upgrades to 92%) does best at matching the actual performance results
	RED	indicates a matchup where the current 400-point rule (maximum of 1 upgrade to 92%) from 2022 does best at matching the actual performance results
	YELLOW	indicates a matchup where completely removing the 400-point-rule does best at matching the actual performance results

Amazingly, there are no other cells that get color-coded as yellow (preferring no 400-point-rule) or red (preferring the "max-1-upgrade" rule). Everything else is blue. When the higher-rated player has been in the range of 2000 to 2399 Elo, or lower than 2000 Elo, it's always been true that the best option was to use the traditional 400-point-rule, with unlimited upgrades to 92%. The average score achieved by these rating favorites has always been 92% or lower in these Extreme Elo Difference games. That's why all these cells are color-coded blue, to indicate that for those ranges of player, the best option would be to go with the traditional 400-point-rule.

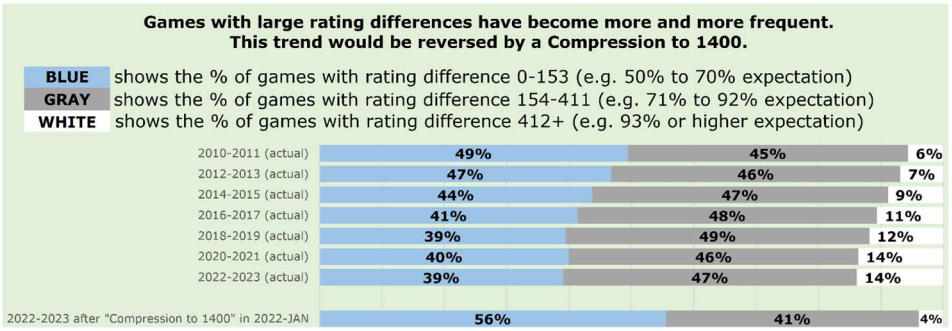
And if we look at the Totals, down at the bottom of each column, we see that the recommendation throughout these years, based on overall averages across all Extreme-Elo-Difference games, would be to go with the traditional 400-point-rule, even taking the games played by IM's and GM's into consideration.

So it would normally be a very easy recommendation, to "follow the blue" and say that as far as these numbers are concerned, we should revert the 400-point rule to its pre-2022 state, going back to the traditional 400-point-rule.

However, I am also advocating a compression together of the ratings, and if we do that, it does change the situation quite notably for the 400-point-rule.

First of all, in the timeframe since 2010, the bottom of the rating list has filled out with lots of players, and so it shouldn't be surprising to hear that Extreme-Elo-Difference games are more frequent now than they used to be. In the following bar charts, we see how the percentage of games that are Extreme-Elo-Difference has steadily increased in the past dozen years:

The white bars over on the right tell us what fraction of games qualified, during each two-year-stretch, for the 400-point-rule. In fact, from 2020-2023 we see that 14% of games have had Elo differences so large that they qualified for the rule.



However, if we do a compression, then ratings will move closer together, and so we won't have nearly so many games with Extreme Elo Differences. In fact, it would drop down to only 4% of all games, as we can see from the very bottom row there.

With all these ratings much closer together, would it impact the choice of which is the ideal version of the 400-point-rule? I checked this by running a simulation with a compression happening two years ago, in January of 2022, and then seeing what would have happened to ratings up to the present, especially what would happen in those 4% of games that would still qualify for the 400-point-rule. This allows us to add a fourth column to our previous picture, over on the far right of the following image:

Extreme-Elo-Difference games (those having Elo difference of 412+)

	Years 2010 and 2011	Years 2016 and 2017	Years 2022 and 2023	Years 2022 and 2023 (after simulated compression in 2022-Jan)
When higher-rated player is rated below 2000 Elo	11,849 games (16% of total) actual: 88.9% score **traditional rule: 92% expect** 1-upgrade rule: 92.3% expect no rule: 95% expect	138,849 games (53% of total) actual: 88.0% score **traditional rule: 92% expect** 1-upgrade rule: 93% expect no rule: 95.9% expect	114,325 games (61% of total) actual: 83.2% score **traditional rule: 92% expect** 1-upgrade rule: 93.3% expect no rule: 96.1% expect	11,013 games (21% of total) actual: 90.9% score **traditional rule: 92% expect** 1-upgrade rule: 92.3% expect no rule: 94.3% expect
When higher-rated player is rated 2000-2399 Elo	43,691 games (60% of total) actual: 92% score **traditional rule: 92% expect** 1-upgrade rule: 92.9% expect no rule: 95.7% expect	99,510 games (38% of total) actual: 91% score **traditional rule: 92% expect** 1-upgrade rule: 93.6% expect no rule: 96.4% expect	63,172 games (34% of total) actual: 87.1% score **traditional rule: 92% expect** 1-upgrade rule: 94.2% expect no rule: 96.9% expect	31,071 games (61% of total) actual: 91.3% score **traditional rule: 92% expect** 1-upgrade rule: 93.3% expect no rule: 95.7% expect
When higher-rated player is rated 2400-2799 Elo	17,090 games (24% of total) actual: 95.4% score traditional rule: 92% expect 1-upgrade rule: 93.3% expect **no rule: 96.1% expect**	21,549 games (8% of total) actual: 94.2% score traditional rule: 92% expect **1-upgrade rule: 93.8% expect** no rule: 96.5% expect	10,259 games (5% of total) actual: 92.1% score **traditional rule: 92% expect** 1-upgrade rule: 94.2% expect no rule: 96.8% expect	9,170 games (18% of total) actual: 92.7% score **traditional rule: 92% expect** 1-upgrade rule: 93.8% expect no rule: 96.2% expect
Total	72,630 games total actual: 92.3% score **traditional rule: 92% expect** 1-upgrade rule: 92.9% expect no rule: 95.7% expect	259,908 games total actual: 89.7% score **traditional rule: 92% expect** 1-upgrade rule: 93.3% expect no rule: 96.1% expect	187,756 games total actual: 85% score **traditional rule: 92% expect** 1-upgrade rule: 93.6% expect no rule: 96.4% expect	51,254 games total actual: 91.5% score **traditional rule: 92% expect** 1-upgrade rule: 93.2% expect no rule: 95.5% expect

LEGEND:		
	BLUE	indicates a matchup where the traditional 400-point-rule (unlimited upgrades to 92%) does best at matching the actual performance results
	RED	indicates a matchup where the current 400-point rule (maximum of 1 upgrade to 92%) from 2022 does best at matching the actual performance results
	YELLOW	indicates a matchup where completely removing the 400-point-rule does best at matching the actual performance results

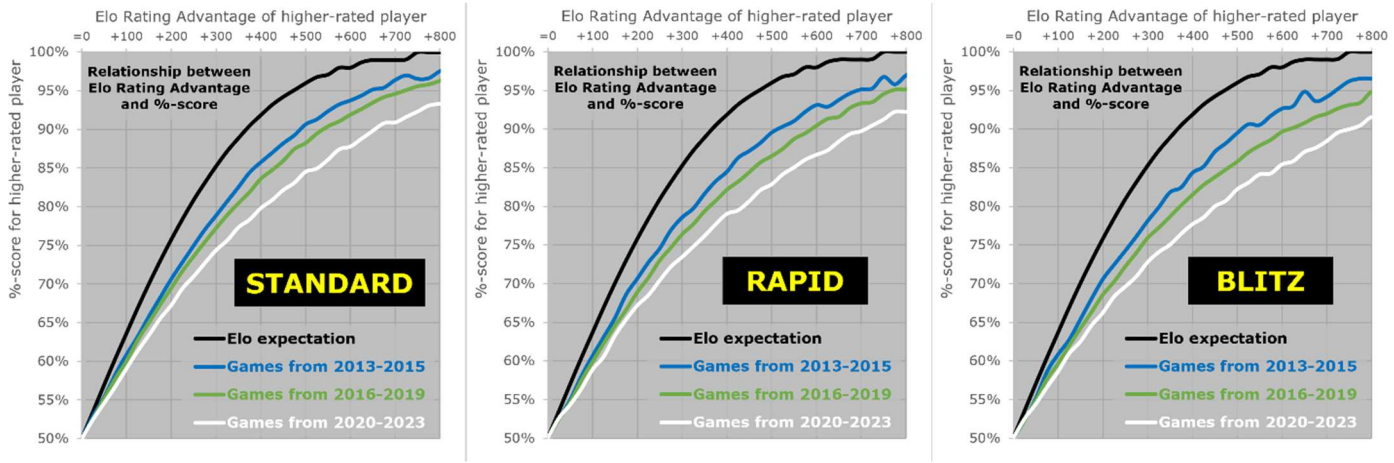
We can see the color-coding for this simulated scenario over on the right there. It still is blue for all ranges of player rating, indicating that even if a compression were to happen, the best choice (for all strengths of the rating favorite) would still be the traditional 400-point-rule, rather than the current "max-1-upgrade" rule or just eliminating any such 400-point-rule.

So I am quite comfortable recommending that based on these numbers, we should revert the 400-point-rule back to its pre-2022 state.

Let me also briefly talk about rapid and blitz as well. The bulk of my analysis pertained to the standard rating system, but as far as I can tell, everything that I've said about standard Elo ratings will apply equally well to rapid and blitz Elo ratings. For the most part, these three rating

systems do not affect each other, but there is some overlap among them, such as when you enter a rapid/blitz event when you have no applicable rating but you do have a standard rating. Essentially it turns out that if we are ever going to ever do the compression and calculation improvements for rapid and blitz too, it works much better if all three rating systems (standard, rapid, and blitz) are affected at the same time. So... would that be a good idea?

We have already looked at the black and white curves showing the relationship between Elo rating advantage and percentage score, but so far in this article, we've only looked at it for Standard rated games. Let's look at rapid and blitz rated games as well, in a side-by-side view:



As you can see from these three graphs, we have seen analogous behavior across all three systems, where there was already a problem eight or ten years ago (as shown by the blue curves), then it got worse over time (as shown by the green curves), and is quite problematic when you look at the white curves representing the last few years. The red and blue "heat maps" that were provided in my original proposal in July, tell much the same story as standard if you produce them for rapid and blitz as well, although those are not shown here.

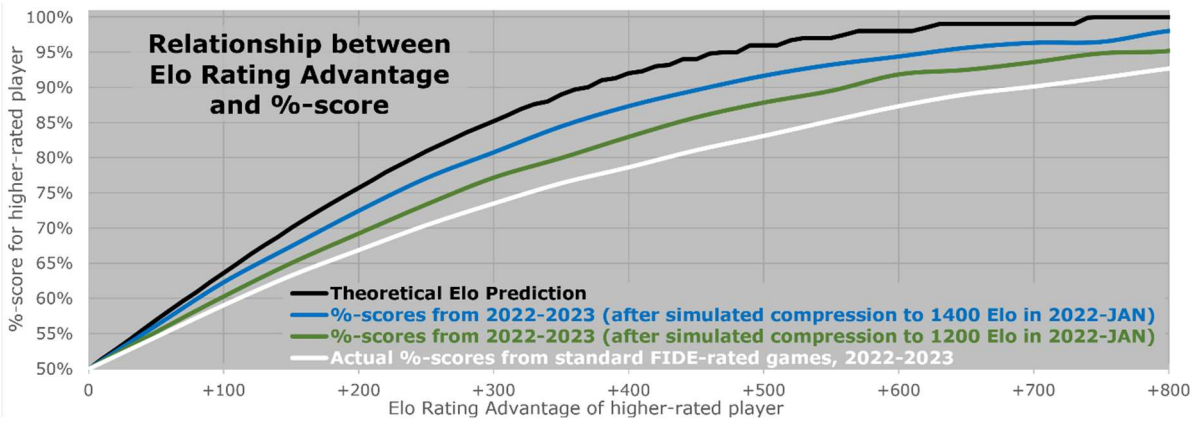
So this deflation (downward rating expansion) seems to be happening just as much in rapid and blitz as it is in standard. My recommendation was that we roll out these exact same changes to rapid and blitz at the same time that we do them for standard.

I also want to call attention to one small difference across the three rating systems. If you look closely, you can see that the white line is even shallower for rapid than it is for standard, and the white blitz line is a bit shallower still. For example, at a 500-point-rating advantage, the rating favorites are scoring about 85% in standard chess, compared to 83% in rapid and 82% in blitz. To me this says that if we are doing a compression for rapid and blitz, then it's even more important to revert the 400-point-rule, because the chaos of rapid and blitz chess makes it even harder for the extreme rating favorite to even approach scoring 92%. So we shouldn't make it extra-hard for them, by retaining the "max-1-upgrade" rule.

So overall, where do I think this will take us? Of course, nobody can predict the future with perfect accuracy. However, we can use simulations to explore the quite pertinent question of, if we had implemented the Compression and Calculation Improvements a couple of years ago, in January of 2022, would we indeed be in a better place today?

We can see how players would have done, relative to their Elo predictions, in the various simulations, if all of the chess games that subsequently were played, had indeed been played (with the same outcomes) during the simulation, using the different ratings. And so this gives us a powerful way to project what might happen to the rating system if we do indeed make these changes. Note that these simulations are just for the Standard rating system.

We saw previously that the actual performance by rating favorites in the last couple of years, fell quite short of the Elo expectation taken from the tables in the FIDE Rating Handbook. As usual, the expected score is shown in black and the actual score in real life is shown in white. But in the following graph, we can also add in the average scores by rating favorites during the simulations. Those are shown in the blue and green colors in the following graphic:



The blue curve assumes we did the full compression (bringing 1000 Elo players to 1400, etc.) in January 2022, and the green curve assumes we did a halfway compression instead (bringing 1000 Elo players to 1200, etc.) And the blue and green curves show players' scores in games during 2022 and 2023 using their simulated (compressed) ratings.

As you can see, these scenarios will not completely remove all traces of the deflation, the main reason being that we are not directly changing the ratings of players rated 2000+ Elo. Nevertheless, even the halfway compression to 1200 Elo minimum would take care of about one-third of the problem, and the full compression to 1400 Elo minimum would evidently take care of about two-thirds of the problem! And I think there's reason to expect this is what would happen in real life as well. So I still feel strongly that it would be very beneficial for the rating systems for us to go ahead and make these changes, as soon as possible.

Finally, let me take the opportunity to thank everyone from the chess community who took the time to prepare and submit feedback on the original July proposals. There were more than 150 responses, and of course no solution and no follow-up-analysis would be able to address everyone's concerns and suggestions. I don't have infinite amounts of time to spend on all this, as much as I might wish that I did. I did try to address several important responses, within my 47-page supplemental report that I prepared for FIDE in late October. So if I don't seem to have addressed your concerns, please consult that supplemental report. If nothing else, please know that the QC members and I did read through and consider all of the submitted feedback.

It is quite possible that FIDE will need to take further steps to combat the ongoing deflation, if these measures prove inadequate. I believe that these measures have preserved the simplicity of calculation that is a hallmark of the Elo system, and I hope and expect them to greatly lessen (or even halt) the ongoing deflationary effect that has dominated the rating systems for the past several years. It is quite a balancing act, and hopefully we have found a good balance.

Although I am an external consultant and I certainly don't speak for the QC, it is my understanding and expectation that the QC will keep monitoring the situation, and will keep an open mind regarding many of the excellent suggestions that were already made.

Thanks for reading all of my long articles!
-- Jeff