# Comments to: Repairing the FIDE Standard Elo rating system.

Otto Milvang, Norwegian chess federation, 18. sept. 2023,

## Summary

The paper discussed the current situation and the consequences of the QC proposal. The rating compression corrects rating deflation over the last 10 years, but it does not solve underlying problems with rating reliability. The problem is complex, and this paper point out some problems: underrated players, geographical differences and rating floor.

It's clear from Sonas' thorough analyses that something has to be done, so the question is if QC proposal on rating compression is the correct tool. This paper shows that software compression does not solve the underlaying reason for rating deflation. A rating floor on 1400 divide the players in rated and unrated players, and in the range 1400-1600 the rating is highly unreliable.

A Deflation Index DI is defined. A compression as proposes by QC, reset the DI to zero, however simulations shows that the DI still increases the following years. The DI will only decrease over years if rating points are inserted into the rating system.

The paper discusses different methods for adding rating points into the rating system per played game, and simulations compares this method versus compression.

The paper also shows that the rating floor is a border that creates unreliable rating for the lowest rated players. It also destroys a natural rating distributed from the players. In the proposed model rating floor is removed, and the simulation shows that this only has advantages.

## The QC proposal and consequences

### Elo as a zero-sum system

The rating system we are using is a in its nature a zero-sum system. This means that it two players Ann rated 1600, and Bob rated 1300 meet, and both have k=20, then the rating changes are:

| Result | Ann:1900 ΔR | Bob:1600 ΔR |
|---|---|---|
| 1-0 | 3 | -3 |
| ½ - ½ | -7 | 7 |
| 0-1 | -17 | 17 |

Suppose that Ann and Bob move to an isolated island and continue to play rated games against each other. Suppose that Ann has a constant plying strength at 1600, while Bob increases his playing strength to 1900. Since the rating system is a zero-sum system, it will preserve the correct rating difference, while the mean level is constant.

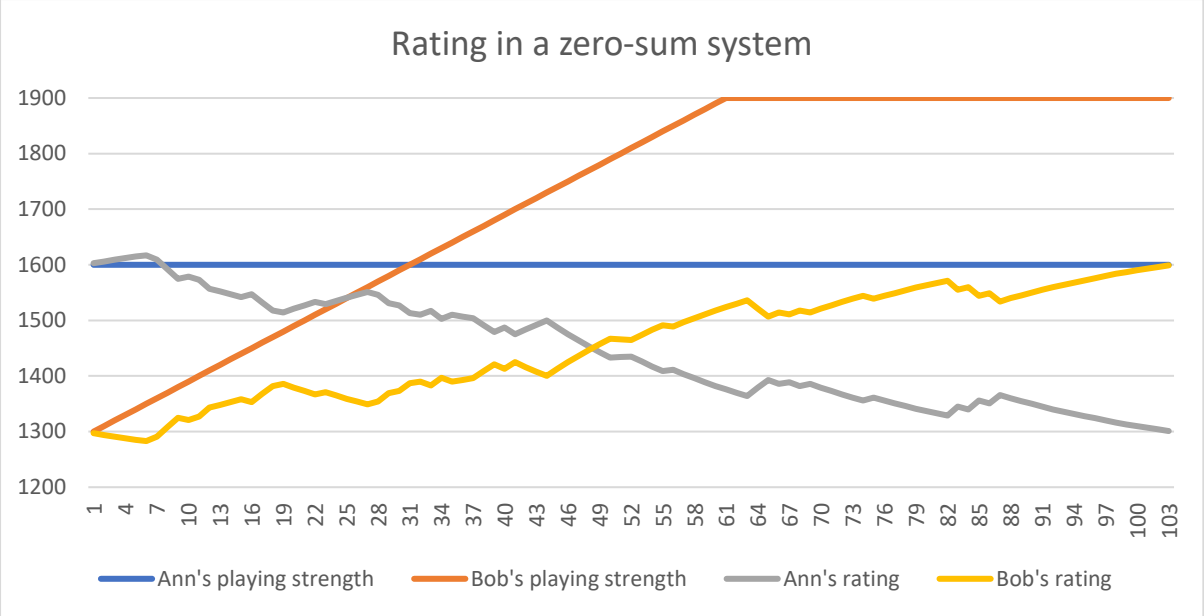



*Figure 1 Simulation of rating development, 103 games,*

As a result, over many games, Ann will end up with ~1300 rating and Bob with ~1600 rating. The difference in rating is correct, while the level is too low.

For young players, and new players FIDE has K=40. This is intended to prevent established players from losing rating at the expense of young and up-and-coming players. Does it work?

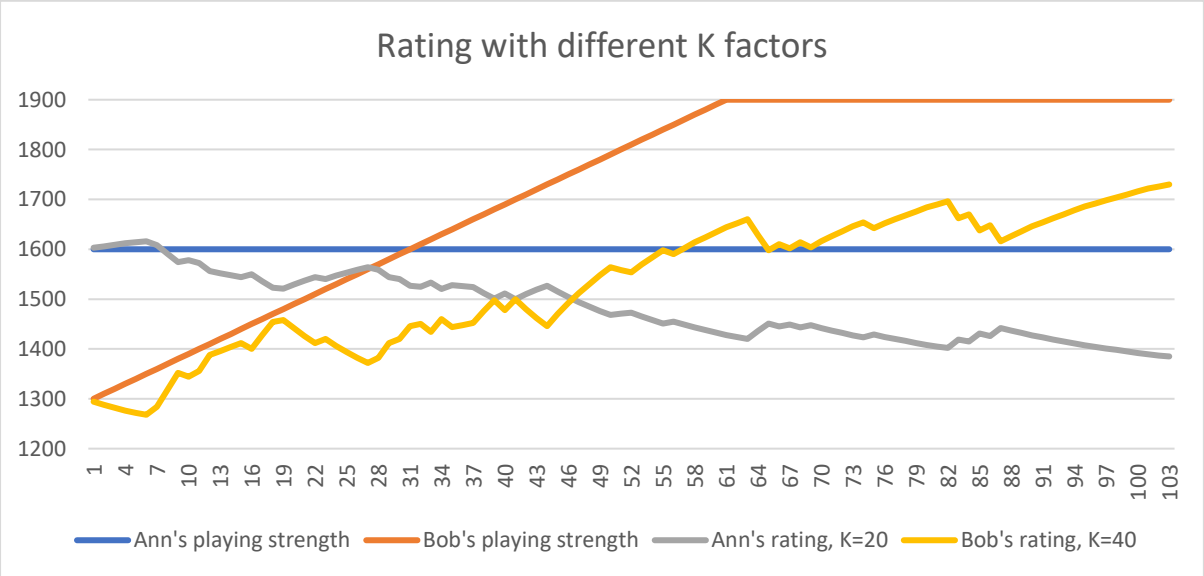



*Figure 2 Simulation of rating development, different K-factors*

Yes, it works, but less than expected. If Ann has K=20, and Bob has K=40 it will lift the rating level only 100 rating points, 300 were needed to maintain a rating system without deflation.

Will a rating compression help? No, not at all!

We will still have ratings in a zero-sum system, the only differences in the simulations are that the x-axis is compressed!

## Strong young players are underrated

A young talented Norwegian chess player, born 2009, played this tournament in February 2022, and his rating before the tournament is 1194. In July 2023 his rating is 1844.

**Standard Ratings March 2022** Total change: **113.60**

| **Tournament in 2022** | | | Oslo | | NOR | 2022-02-18 | | |
|---|---|---|---|---|---|---|---|---|
| **Rc** | **Ro** | | | **W** | **n** | **Chg** | **K** | **K*chg** |
| 1773 | 1194 | | | 3.00 | 5 | 2.84 | 40 | 113.60 |
| ■ B, I | | 1853 | NOR | 0.00 | 1 | -0.01 | 40 | -0.40 |
| □ K, E | | 1588 | NOR | 1.00 | 1 | 0.92 | 40 | 36.80 |
| ■ H, L | | 1710 | NOR | 1.00 | 1 | 0.96 | 40 | 38.40 |
| □ L, | | 1750 | SWE | 1.00 | 1 | 0.97 | 40 | 38.80 |
| ■ A, E | | 1964 | NOR | 0.00 | 1 | 0.00 | 40 | 0.00 |

With the rating compression proposed by QC, the same tournament with rating adjustment.

**Standard Ratings March 2024** Total change: **96.40**

| **Compressed rating 2024** | | | Oslo | | NOR | 2024-02-18 | | |
|---|---|---|---|---|---|---|---|---|
| **Rc** | **Ro** | | | **W** | **n** | **Chg** | **K** | **K*chg** |
| 1864 | 1516 | | | 3.00 | 5 | 2.41 | 40 | 96.40 |
| ■ B, I | | 1912 | NOR | 0.00 | 1 | -0.08 | 40 | -3.20 |
| □ K, E | | 1753 | NOR | 1.00 | 1 | 0.80 | 40 | 32.00 |
| ■ H, L | | 1826 | NOR | 1.00 | 1 | 0.86 | 40 | 34.40 |
| □ L, V | | 1850 | SWE | 1.00 | 1 | 0.88 | 40 | 35.20 |
| ■ A, E | | 1978 | NOR | 0.00 | 1 | -0.05 | 40 | -2.00 |

The fact that many young players are underrated is the challenge. Underrated players decrease the rating of established players (players with K=20). The proposal from QC does not solve this problem. We will still have talented young players that climb up the rating at the expense of established players. With QC-proposal, the problem is even worse if the K-factors are unchanged since Chg=2.41 with compressed rating is equal to Chg=4.02 in our current rating system.

## Cultures

Rating is in many cases closed ecosystems, as evolve different in different countries.

If we look at rating in the countries with most active players (played as least one game after JAN21) we got:

| FED | Count | Mean | q10 | q90 |
|---|---|---|---|---|
| ESP | 17687 | **1601** | 2028 | 1194 |
| FRA | 15421 | **1514** | 1976 | 1125 |
| GER | 12107 | **1803** | 2176 | 1402 |
| IND | 11566 | **1298** | 1675 | 1050 |
| RUS | 10536 | **1487** | 2090 | 1074 |
| ITA | 6556 | **1574** | 2001 | 1180 |
| POL | 6282 | **1483** | 2009 | 1094 |
| CZE | 5830 | **1728** | 2095 | 1305 |
| IRI | 4769 | **1417** | 1835 | 1092 |
| USA | 4149 | **1753** | 2134 | 1393 |
| TUR | 4121 | **1412** | 1861 | 1089 |

*Table 1 Mean FIDE rating for the countries with most active players. Count is the number of players, q10 and q90 are the 10% and 90% quantiles.*

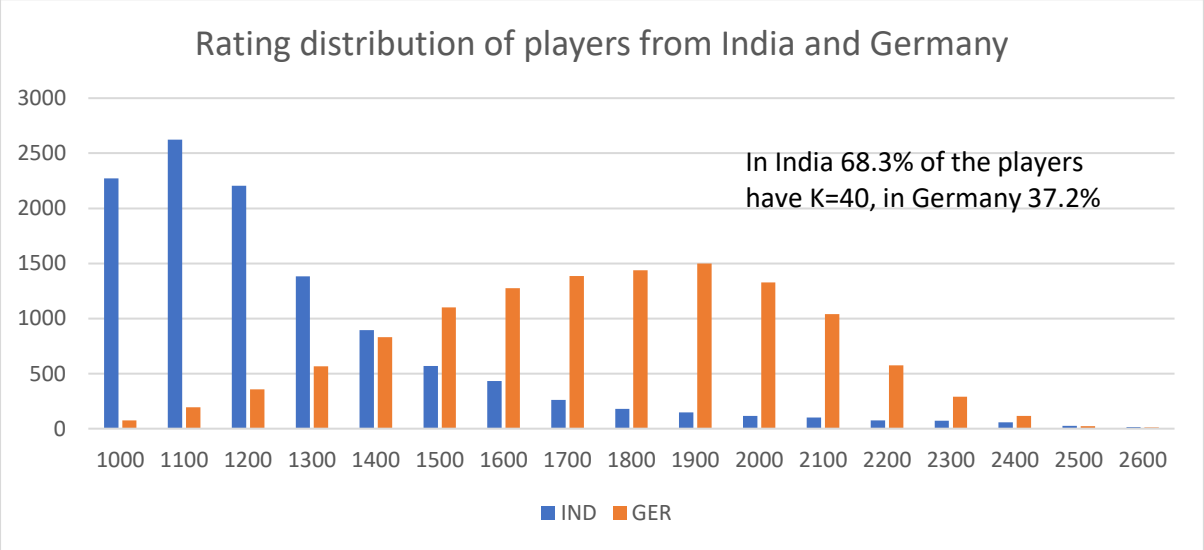



*Figure 3 The number of players in steps of 100 ration points.*

Table 1 and Figure 3 shows a huge difference in rating for players in Germany and India. It is impossible to say if this different is real or artificial.

Will a rating compression help? No!

## Rating floor

After the rating floor was set to 1000 in 2012 the number of players with FIDE rating has increased every year.

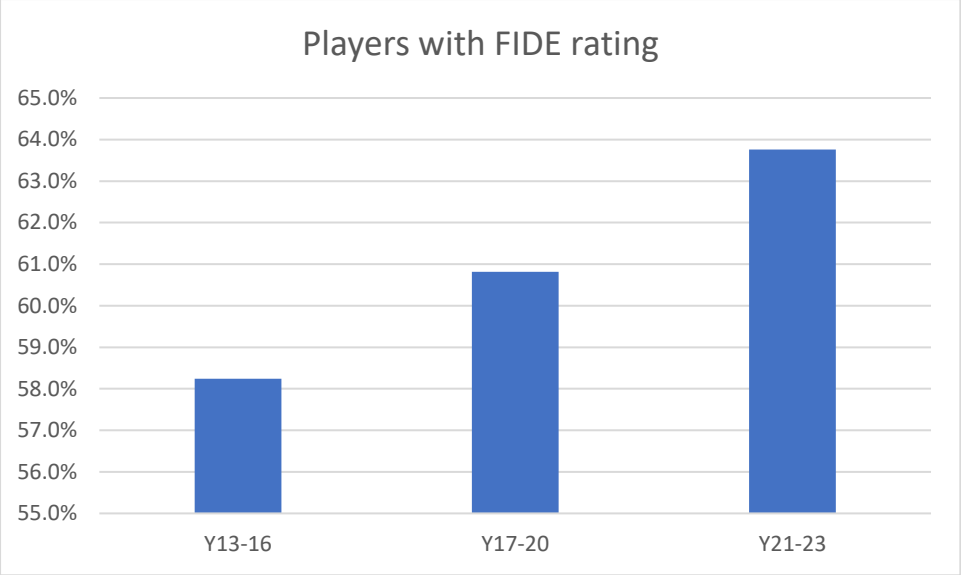



*Figure 4 Players with FIDE rating that has played within different time periods.*

Figure 4 shows that the number of FIDE rated players are increasing, and many federations thought FIDEs plan was to decrease rating floor such that almost all players could have FIDE rating.

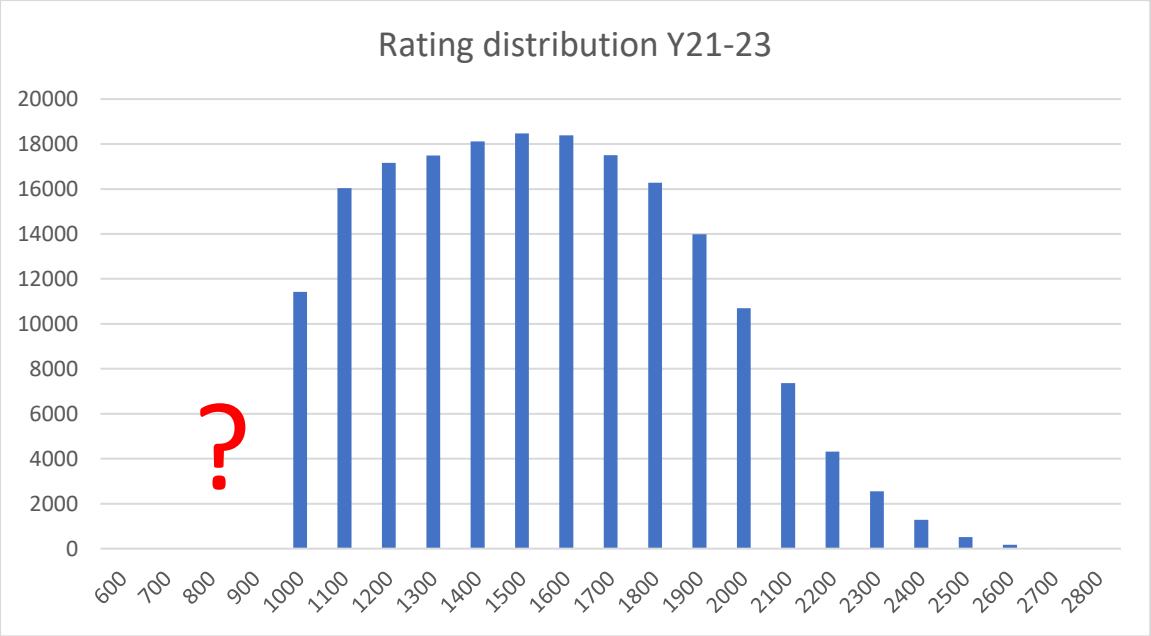



*Figure 5 Rating distribution, players played at least one game 2021-2023*

Figure 5 shows the current rating distribution of active players. It's clear that in a distribution of players something is missing. Sonas claims that decreasing the rating floor will start pulling large amounts of rating points away from the established pool. This is true also for the QC proposal, so the problem is not that low rated players increase their playing skills, but that there are no mechanisms to prevent deflation.

**The chess community expected FIDE to decrease the rating floor such that the rating system would be a consistent system for all chess players.** The Norwegian chess federation had a rating system down to 600 until 2017, where it was supposed that FIDE rating would be the main rating system. Already in 2017 it was clear that the fact that it is a rating floor with 2/3 of the players with rating

and 1/3 without rating is challenging. The floor gives high degree of uncertainty within the first 300 rating points (1000-1300).

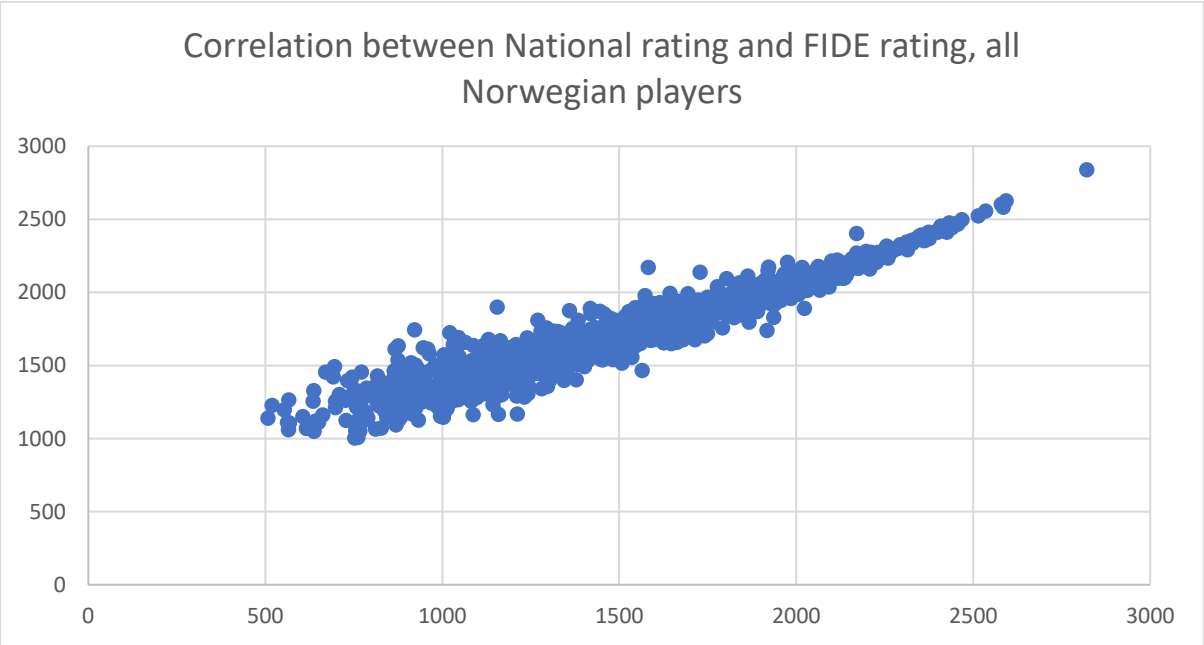



*Figure 6 Correlation between National rating and FIDE rating*

The National rating worked well. The players with local rating 600-1000 had FIDE rating in the range 1000-1500, which is a huge span. The reason is that the initial rating is more or less random.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ■ S, D | | | 1193 | NOR | 0.00 | 1 | 0.00 | 0.00 |
| ■ M, L | | | 1238 | NOR | 0.00 | 1 | 0.00 | 0.00 |
| □ F, S | | | 1261 | NOR | 0.50 | 1 | 0.00 | 0.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| □ H, O | | | 2119 | NOR | 0.00 | 1 | 0.00 | 0.00 |
| □ H, P. | | | 1868 | NOR | 0.00 | 1 | 0.00 | 0.00 |
| ■ N, O | | | 1592 | NOR | 0.00 | 1 | 0.00 | 0.00 |
| □ F, G | | | 1743 | NOR | 0.00 | 1 | 0.00 | 0.00 |

In this example a player that got his initial rating 1152 after score 0.5 / 7.0 against a mix of weak and strong players. With established rating on 950, k=20 his rating would be 944. This shows that initial rating is highly unprecise, and rating should be introduced on a lower rating level.

This is one of many examples where new players have an artificial high rating compared to their playing strength. There are also players that are underrated in the same rating range. As a result of a rating floor that divide the players in rated and unrated players. As a result, you cannot guess anything about the playing strength of a 1200 rated player compared to a 1000 rated player.

Will a rating compression help? No!

As long as there is a rating floor dividing the players in rated and unrated players there will be turbulences in the border area. We can expect this uncertainty to cover the same rating spam as today (since this is the nature of rating.

With the example above, the initial rating would with two extra draws against 1800, be 1.5p on [1800,1800, 1516, 1543, 1557, 2119, 1921, 1755, 1846]) => 1489. This rating is at least 1000 rating points too high and makes the rating for the lower part of the scale unreliable.

Why does FIDE still have a rating floor?

No other chess playing platforms have rating floor. New players expect to have rating from game 1. The quality of the Swiss paring is based on the quality of the rating. To remove the rating floor will also improve the pairing.

### Existing tournament formats

Several chess events are based on the current rating model with typical rating limits 1200,1400,1600, and so on. This is well incorporated, and after the rating compression, either the rating limits will be artificial, or completely new models must be developed for the events.

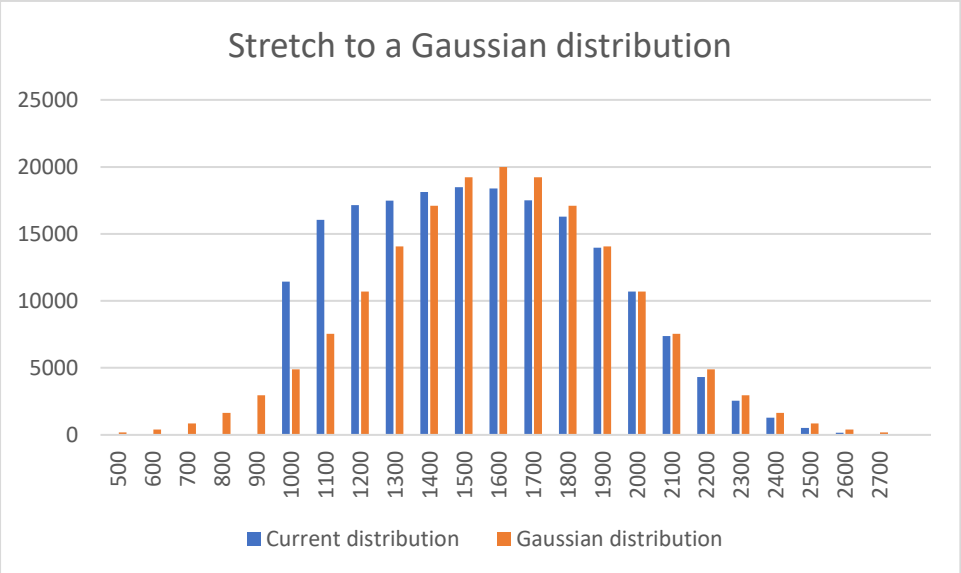
## How to solve the deflation problem

### Linear stretching/compression

The QC proposal.

### Gaussian stretching/compression

Instead of compressing the rating distribution, the rating distribution can be stretch to a Gaussian distribution.




This is an interesting approach but requires that all players have a rating or a provisional rating. It still has weaknesses since we do not know if playing strength is Gaussian distributed.

## Adjust the formula for rating calculation.

The goal is to insert rating point into the rating system, meanwhile the rating system is a zero-sum system with equal k-factors. As shown in the previous section different k-factors does not insert enough rating point to prevent deflation. An approach is it modify the formula for scoring probability PD.

$$PD = RDtoPD(Rc - Rp)$$

Where Rc is opponent rating, Rp is player rating and RDtoPD is given by table 8.1.2 in FIDE rating regulations. This formula works well if both players rating reflect their playing strength. The problem is that many young players, and new players are underrated due to fast progress in playing strength. A solution is to add extra rating into the formula if the opponent is young or has few rated games.

**Proposal 1:**

**In rating calculations, when the opponent has kFactor = 40 and rating Rc below a limit L, adjust PD**

$$PD = RDtoPD(Rc + (L-Rc)/1000 * \lambda + -Rp)$$

**where typical values for L and λ can be L = 2000 and λ = 80**

## Add rating to games played

Another way to insert rating point into the rating system is to add an amount of rating point for each played games by players who is a phase where they gain many rating points. This is easy to add by adding this to ΔR in

$$\Delta R = score - PD + \alpha$$

Again, the problem is that many young players, and new players are underrated due to fast progress in playing strength.

**Proposal 2:**

**In rating calculations, when the player has kFactor = 40 and rating Rc below a limit L, compute ΔR by**

$$\Delta R = score - PD + (L-Rr)/1000 * \beta$$

**where typical values for L and λ can be L = 2000 and β = 2.0/k**

## Remove the rating floor

With proposal 1, and 2, there are no reason to keep the current rating floor. Players expect to have rating from game 1, and when rating points are inserted to avoid deflation there are no reason to keep this artificial limit between rated and unrated players.

**Proposal 3:**

**A player new to the rating system is given 1500 as a provisional rating for adults, and for children 1200 until the end of the year of their 18th birthday. The provisional rating for the next 15 games is calculated as the rating performance of the games played + 2 x draw against their initial rating (1500 / 1200). After the month where at least 15 games against rated players are played the player will receive a normal rating set to his provisional rating. Games against players with provisional rating is not rated.**

**Adult players shall have K=40 for the next 30 games.**

## How to measure deflation, define Deflation index

Sonas has described the deflation with tables and graphs like:

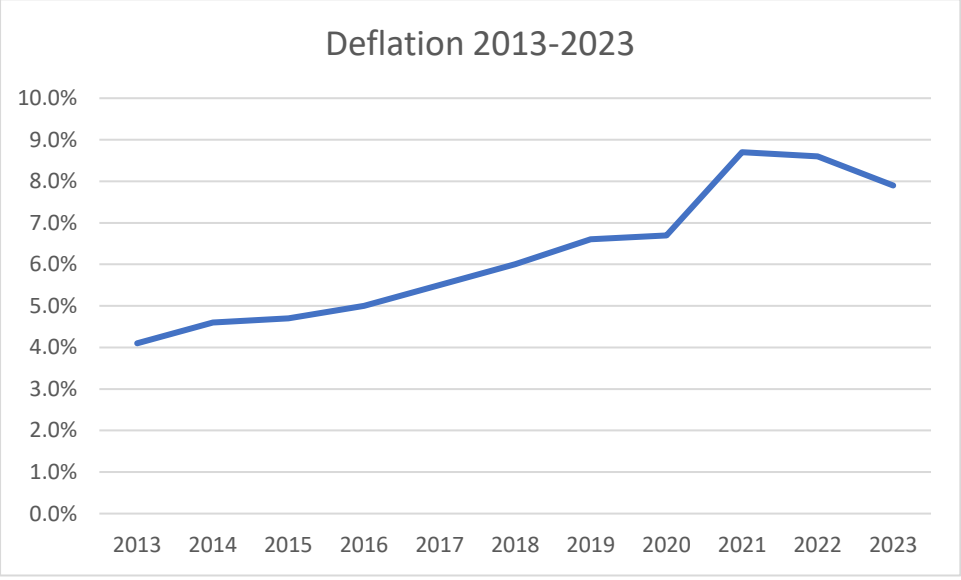
| 2021-2023 | opponent / player | vs. 1000-99 | vs. 1100-99 | vs. 1200-99 | vs. 1300-99 | vs. 1400-99 | vs. 1500-99 | vs. 1600-99 | vs. 1700-99 | vs. 1800-99 | vs. 1900-99 | vs. 2000-99 | vs. 2100-99 | vs. 2200-99 | vs. 2300-99 | vs. 2400-99 | vs. 2500+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| +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% | |

This table (from Sonas' paper) describe the difference between actual score and expected score. This means that in the current rating system a higher rated player scores less than expected against lower rated players.

Its important to know how deflation evolve, so a Deflation Index DI is defined as

DI = - Sum of all games(actual score for **higher rated player** – expected score)/ number of games.

Note the minus sign before the summation to get a positive DI value. DI is the mean of the values in upper right triangle adjusted for number of games. For 2021-2023 DI = 8.5%




The graph shows that the Corona years 2021-2022 are untypical, and if may be unwise to draw conclusion on games in this period.

## Simulation

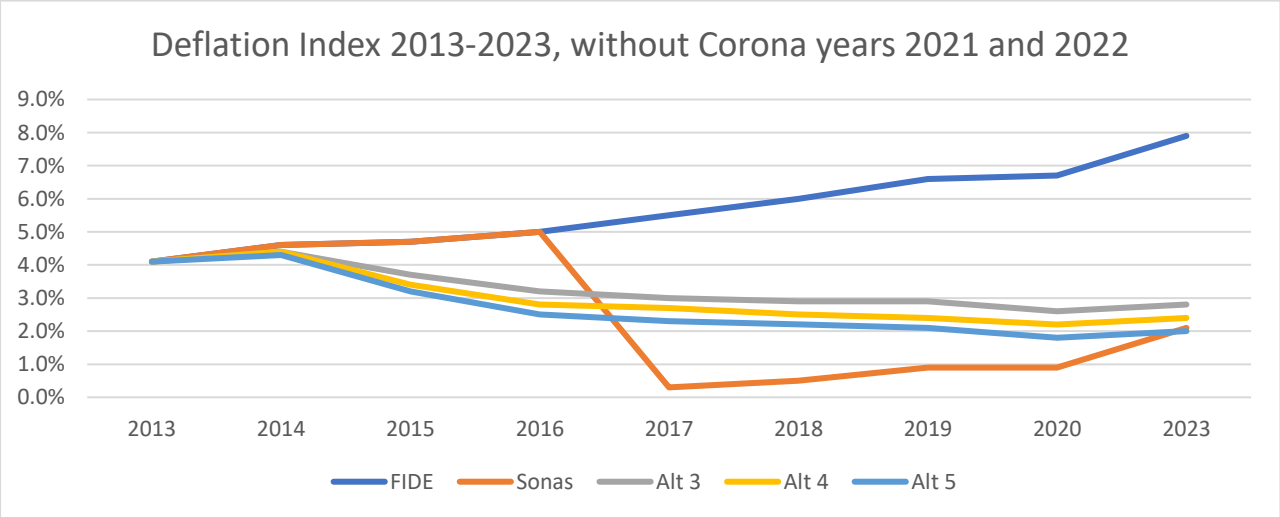
All the simulations are done on the same database as Sonas' used. Only games with standard rating are used in the simulations.

A huge set of parameters were tested, and only a few are selected for presentation to show the effect of the measures. The models are:

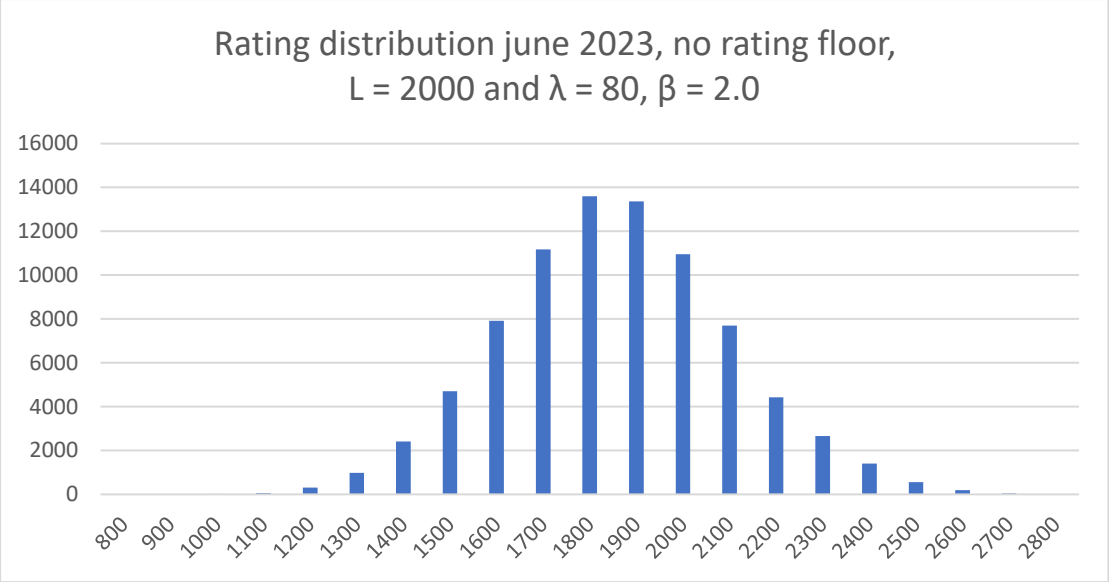
1) No change (FIDE today)
2) The Sonas proposal, Adjust and start JAN 2017
3) L = 2000 and λ = 80, β = 2.0, Start JAN 2014
4) L = 2000 and λ = 100, β = 3.0, Start JAN 2014
5) L = 2000 and λ = 120, β = 4.0, Start JAN 2014

### Result of simulation:

| Year | FIDE | Sonas | Alt 3 | Alt 4 | Alt 5 |
|---|---|---|---|---|---|
| 2013 | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% |
| 2014 | 4.6% | 4.6% | 4.4% | 4.4% | 4.3% |
| 2015 | 4.7% | 4.7% | 3.7% | 3.4% | 3.2% |
| 2016 | 5.0% | 5.0% | 3.2% | 2.8% | 2.5% |
| 2017 | 5.5% | 0.3% | 3.0% | 2.7% | 2.3% |
| 2018 | 6.0% | 0.5% | 2.9% | 2.5% | 2.2% |
| 2019 | 6.6% | 0.9% | 2.9% | 2.4% | 2.1% |
| 2020 | 6.7% | 0.9% | 2.6% | 2.2% | 1.8% |
| 2021 | 8.7% | 2.9% | 3.5% | 3.0% | 2.7% |
| 2022 | 8.6% | 2.6% | 3.5% | 3.0% | 2.6% |
| 2023 | 7.9% | 2.1% | 2.8% | 2.4% | 2.0% |




The table and graphs show how Deflation n Indexes are evolved with different models. Sonas' model is just an offset at 2017 of the current deflation. Alt 3-5 works in the same way with different slope.




The histogram shows the simulated rating distribution 1. june 2023 for players played at least one game in 2023. This is based on rating simulation since Jan14 with no rating floor, L = 2000 and λ = 80, β = 2.0.

### Discussion:

The simulations shows clearly that the QC proposal doesn't solve the problem, The slope of the long-term deflation is the same as the current deflation. This means that in 10-15 years the deflation will be the same as today. Deflation means that the players in the range 1400-2000 are spread out, means that the weakest players are pushed out of the rating list, and the strongest players are pushed over 2000, which will shift the entire rating scale up.

For the proposed measures to add rating to the players, the slope of the long-term deflation is negative, means that the deflation trend has been reversed. Its important to set a relative low value to λ and β, so it reflects the players strength (maybe λ = 80 is too high).

The rating floor is removed, but the insertion of rating to new players has lifted the entire distribution so few players have rating less than 1000.

## Conclusion

The QC proposal will reset the current deflation in the FIDE rating system, however it will not stop the deflation for the future.

FIDE need to introduce a long-term rating regulation that reverses the deflation, and with a rating system that covers all chess players. This paper has shown methods to reverse the deflation. Parameters can be adjusted to reverse the deflation with different speed. In general changes should as small as possible to avoid damages to the rating system, or the trust on the rating system. L = 2000 and λ = 80, β = 2.0 and no rating floor seems to be good values. With these values the rating system will we totally restored over 20-30 years without any negative consequences.

## Further work

The selection of parameters can be further optimized. One solution to the culture differences is to give different parameters to different countries. Other means can be to adjust rating per country. It's a lot of work to find a good method to equalize countries and beyond the scope of this paper.

## Acknowledgement

Thanks to FIDE for giving me access to the game database, and to Jeff Sonas for his big work for the organization and structuring the data, and for the analyses already presented.

I will also thank all people that have had contributions in the discussion and have commented on my drafts.