QC Annex 3

Report about K factor by Jeff Sonas

In some of my writings this year I have described the significant amount of inflation that can be seen in the FIDE ratings since 1985, especially through 1997. I have also explained why I don't think the inflation came from an overall advancement in chess skill, or from a simple increase in the number of rated players.

However, I have not provided much explanation for where I think the inflation actually did come from. Of course there are many different factors working together in the whole rating system, some inflationary and some deflationary, and it's hard to point the finger at just one of them. It's also very hard to simulate the older workings of the rating system, because we really don't have very good tournament data prior to 1999.

Fortunately, there is a lot of good data available from the past ten years, and I have been able to analyze it in some detail. I would like to summarize this analysis and describe my main theory of why we have continued to see inflation. First, though, it will be useful to briefly review how a new player progresses through the four categories of K-factor.

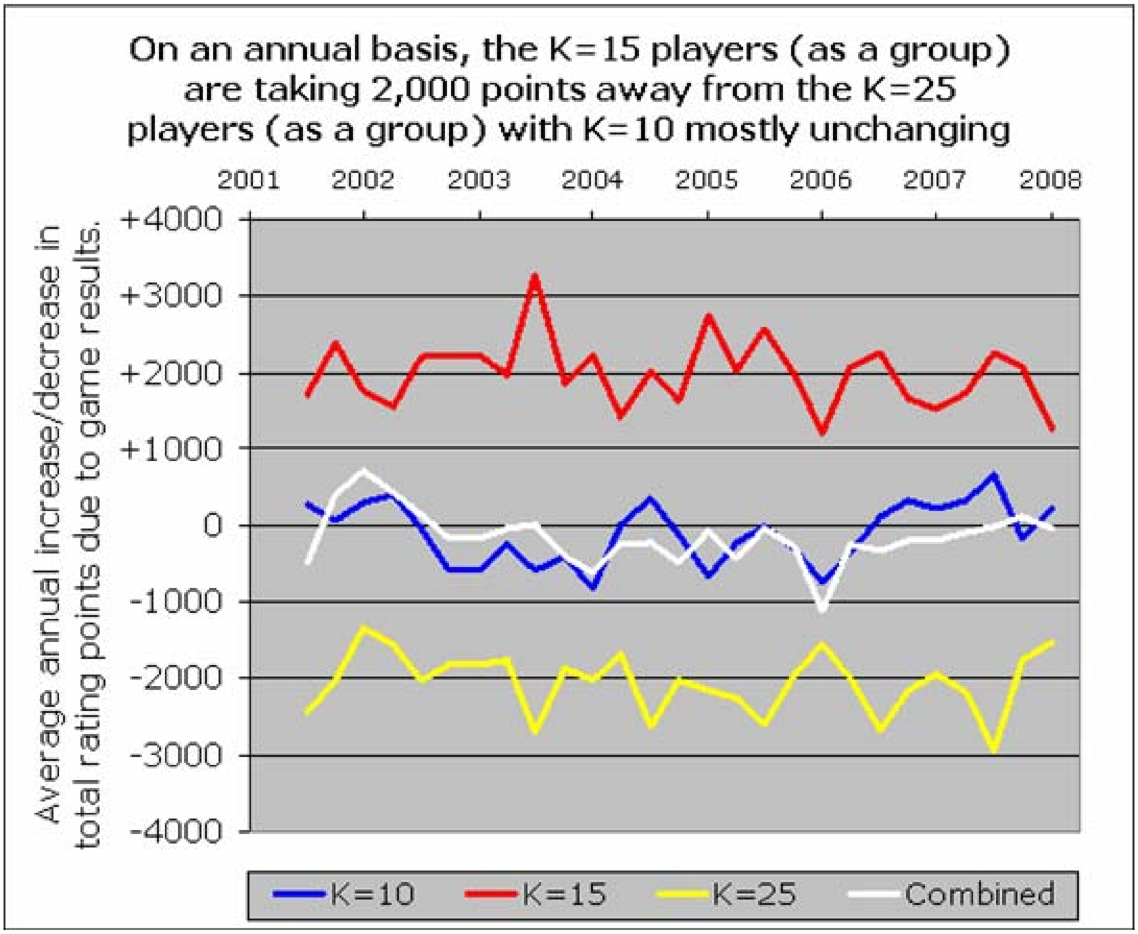
Unrated	Fewer than 9 rated games
K=25	Between 9 and 29 rated games
K=15	30+ rated games, never rated 2400+
K=10	30+ rated games, rated 2400+ at some point

Obviously you start out with no rating. It takes nine games against rated opponents until you can get a rating. Your opponents' ratings are typically not at risk in these games - only your rating is affected - because the rating system ignores games against unrated players (although there are special rules for round-robins with unrated players). It has been suggested that toward the end of a Swiss tournament, rated players who are not in contention might not try their hardest if they find themselves facing an unrated opponent. Keep that in mind, as we work our way through this, because it is actually quite relevant to the inflation question.

Once you have reached nine games against rated players, assuming those games meet a few other conditions, you will get your initial rating on the next list. You will have a K-factor of 25 at that point, meaning your rating will be quite sensitive to wins and losses, making it easier for the rating to "find" your true level. You will keep that K-factor of 25 until you reach a total of 30 rated games. At that point your K-factor goes down to 15, unless you ever reach a rating of 2400+, at which point your K-factor will forever be 10.

Certainly the most intuitive source of inflation comes from the interactions between different K-factors, making it no longer a zero-sum game. If a K=25 player defeats a K=15 player and thereby gains 10 rating points, the K=15 player in turn would only lose 6 rating points. So magically we have just added 4 points to the rating pool, and if this kind of thing happens systematically, then the excess rating points would get distributed throughout lots of players, and ultimately everyone's rating would increase from this. Of course if the reverse happened more than expected, then it would cause deflation rather than inflation.

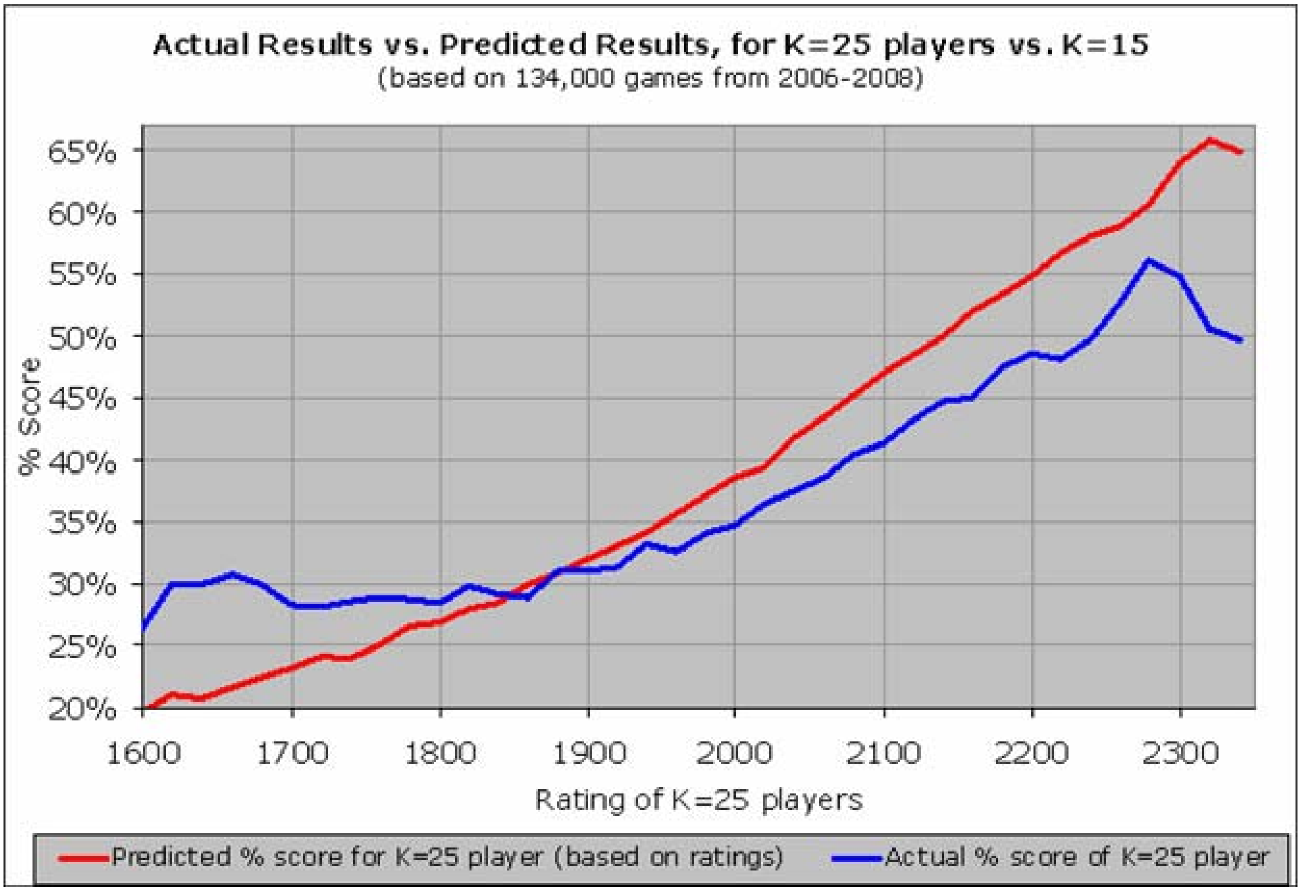
I have data from 2001 to the present, allowing investigation of this as a possible source of inflation. I can tell you that there is no evidence of inflation from this effect. In fact, it's the opposite. The K=25 players do significantly worse than expected against K=15 players, whereas the K=10 players do about as expected against both, and thus the overall impact from this is slightly deflationary, since the K=25 players are losing slightly more than the other players are gaining. Look at the following graph:



The graph reflects the sum of rating changes from all tournament/match results from 2001-2008, shown as an annual average. The red line shows that as a group, K=15 players are overall gaining points from their games; namely, the pool of K=15 players are gaining about 2,000 points per year (as a group). Now, if two K=15 players face each other, then no matter whether the result be a win, loss, or draw, rating points are conserved, and so the entire pool of K=15 players has not gained or lost any rating points. Thus if the whole pool of K=15 players are gaining a total of 2,000 rating points each year, then that can only be at the expense of K=10 and/or K=25 opponents.

Similarly, K=25 (the yellow line) are losing points from their games, which can only be at the expense of K=10 and/or K=15. Since the total change for K=10 (the blue line) is pretty much zero, all of this suggests that K=15 is taking away points from K=25. And if you add up all three groups you get the white line, which is somewhat below zero (which is why I say it is slightly deflationary). Nevertheless this rating point exchange alone is not a major source of inflation, even for the K=15 players, since it would amount to a fraction of a rating point annually per player.

However, this evidence points us in a very important direction. The K=25 players are losing lots of points, consistently, to the K=15 players, their main opponents (K=25 players play about 80% of their games against K=15 opponents). Why are they losing so many rating points? Well, to put it bluntly, the new players (i.e. the K=25 players) seem to be significantly overrated, as a group. Let's look a little closer at the results when K=25 players (of different rating levels) play against K=15 opponents. The following graph shows the predicted %-score for the K=25 players (in red), compared to their actual %-score (in blue):



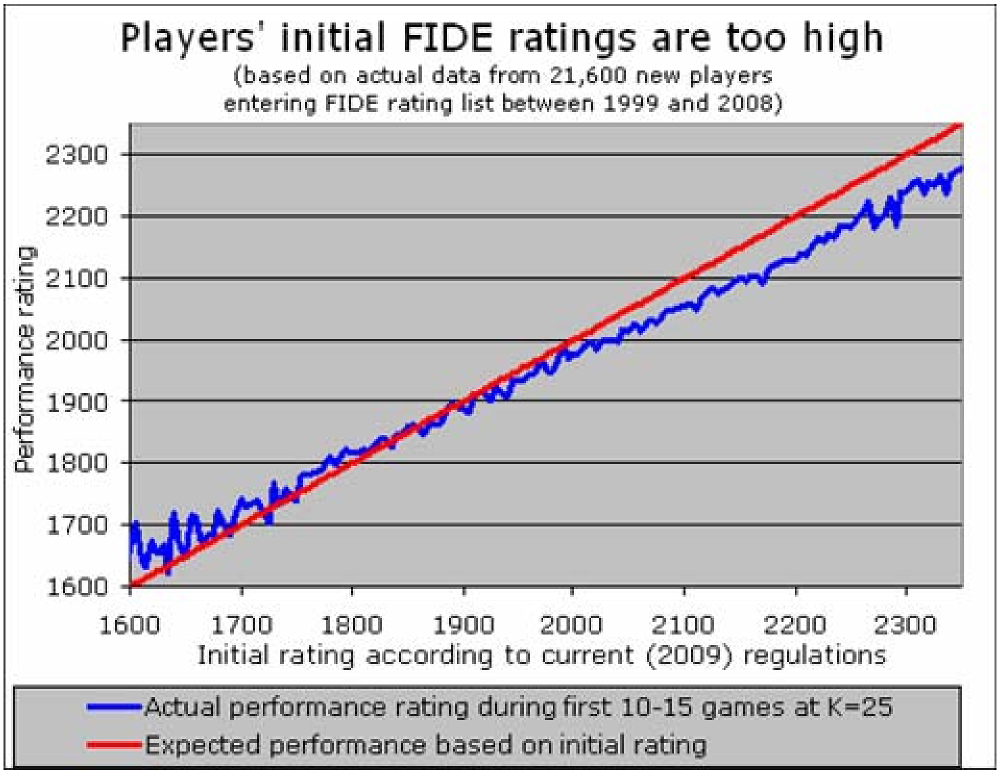
Let me first just make sure you understand this graph. You see down in the bottom-left corner of the graph that the blue line starts at 27%, and the red line starts at 20%. That means that if you look at the results for K=25 players whose rating is around 1600, they are predicted (from the ratings) to score 20%, but they are actually scoring 27%. So lower-rated new players are actually stronger than their ratings would suggest, since the blue line is higher than the red line for players rated below 1800. However, the new players who are rated 2000 or better are definitely under-performing, as you can see by the fact that the blue line is way beneath the red line as you look at the right side of the graph.

What this says to me is that there is way too much variability in initial ratings. The low ones are too low, and the high ones are too high. The initial ratings ought to have much more clustering in the 1800-2000 range. I tried looking at this data a number of ways, also focusing on the K=10 and K=15 groups, and on all the other graphs the red and blue lines were pretty much right on top of each other; it is only K=25 vs. K=15 where we see this anomaly. I am quite certain that the K=25 players are overrated as a group.

I also tried one other approach at investigating this issue. There were some new rules implemented in July 2009 regarding how initial ratings would be calculated, having to do with how the results from your first few events are combined. It used to be that the individual performance ratings from each event were combined, but now the individual games are combined as though they were all in the same event. This is certainly a fairer way of doing things, though it will make inflation even a bit worse. Anyway, if the regulations have changed, then how relevant is the above analysis? Is there any way we can use the new rules but on the old data?

Using the data provided by FIDE from 1999-2008, it was very easy to retroactively calculate in the past what new players' initial ratings would have been (using the rules from the July 2009 regulations) and then to look at how they did over their next couple of events as a K=25 player. This would very clearly indicate whether the current rules for calculating new ratings would indeed show the same behavior as that last graph, where players given low initial ratings do better than expected, and players given high initial ratings do worse than expected.

There were more than 21,000 different players that I could calculate retroactive initial ratings for (on whatever historical rating list where they first reached nine rated games). And for each of those players, I then calculated their performance rating over their next 10-15 games. If their initial ratings were accurate then we would expect that the performance rating would tend to match their initial rating, or maybe even be a little higher, since those players would still be improving. What did I find? Pretty much the same thing that last graph showed.



On this graph, the red line shows you the players' expected performance ratings (i.e. their initial ratings) as well as their actual performance ratings over their first 10-15 games. We see the same pattern as before - players given very low initial ratings will tend to outperform their rating, whereas players given high initial ratings will underperform their rating. More than half of the players in this study were given a (retroactive) initial rating of 2000 or higher, and I would say these players, on average, were overrated by about 60 points.

I believe that this is ultimately the source of the inflation. New players are coming in with ratings that are often too high, and they gradually lose those excess points to the rest of the rating pool, and the points spread out, contributing to inflation. But in the meantime, there are more new players coming in, and they are also facing overrated players. This second wave of new players would already tend to be overrated (as demonstrated above) but in addition, their opponents' ratings are too high, which will inflate the ratings of the new players even more. I think of this as a "compound interest" kind of effect; in the same way you can get additional interest from the interest you've already gained, you can get additional inflation from the inflation we already have.

I have nothing to indicate that this is the only source of inflation. I would expect that if I could run simulations of different formulas for calculating initial ratings, we could start to draw conclusions about whether this is the complete explanation for where inflation comes from, or if it is only part of the puzzle. However I have not yet run these simulations so I am not sure. Nevertheless I think this data is pretty compelling.

So, if you accept my evidence, and my claim that the overrating of brand-new players is ultimately leading to inflation, then what do we do about it? Earlier this fall, I had some very interesting email exchanges with Stewart Reuben and Nick Faulks regarding this question. I find it very interesting that my inclination is quite different from Stewart's, despite the fact that we started from the same data.

First of all, if I may try to characterize Stewart's position, he feels that we can point the finger directly at the fact that established players ease up when facing unrated opponents, unless the game happens to mean something towards the final prizes in a tournament. This effect can make the unrated players' results look more impressive than they really deserve, and ultimately leads to initial ratings being too high, on average. So Stewart feels that we should address this through a rules change whereby games against unrated players are not freebies. This will force established players to try harder against unrated players and hopefully restore the balance. So in his opinion we should implement some sort of rules change to enforce this, and then watch the system for a couple of years and see if initial ratings are still coming in too high, and to what level.

It never even occurred to me to do what Stewart suggested. My approach was that we can leave the entire system exactly as it currently is, with one exception: we change the formula used to calculate initial ratings. Based on fitting various possible formulas against the evidence in that last graph, I came up with a reasonably formula that matches the blue to the red much more closely. If I am right, then this should remove the new inflation, although it might take a couple of years for existing excess rating points to distribute throughout the system.

Stewart feels that it would be impractical to consider such a radical change at this point (i.e. in Halkidiki) because of the lack of advance warning. I can appreciate that, but I do think this approach should be considered at some point, so I am now going to describe what is different about the proposed formula:

In the current formula, if you scored 50% or worse in your initial games, then your new rating is simply your performance rating. But if you had a plus score, then you get your opponents' average rating, plus 12.5 points for every plus you scored. In effect, this means if you had a plus score, then it's as though you started with the same rating as your opponents, and had K=25 already for those initial games, and started gaining 12.5 points with each plus result.

I propose getting rid of the 12.5 point rule, and just using a performance rating no matter if you had a plus score, a minus score, or a 50% score. However, the calculations of this performance rating should be modified in two key ways:

(a) When calculating the initial rating, each player should be treated as though they had an additional 4/10 score among their unrated games, but of course they still need a minimum of 9 real games just as in the current regulations. So the conditions under which you can get an initial rating are unchanged; it is just the calculation of them that is different. Note that in some cases, in fact in many cases, this extra 4/10 will help their percent score, and that is by design, because the initial ratings of players who do quite poorly in their unrated games seem to have been too low.

(b) The opponents are always treated as being 40 points weaker than they really were, for purposes of calculating the performance rating.

So after adding the additional 4/10 score, and after treating the opponents as weaker by 40 points, calculate a performance rating as normal and ALWAYS use that, even for a plus score. You will note in the last example below that even after adding the additional 4/10 score, the player still has a plus score. But the "12.5 per plus" approach is now completely gone; we always use the adjusted performance rating as their initial rating, in all cases.

To illustrate the differences, I picked four examples completely at random:

(A) Unrated player scores 2/9 against avg. 2100 opponents
(B) Unrated player scores 4.5/9 against avg. 1950 opponents
(C) Unrated player scores 5/9 against avg. 2300 opponents
(D) Unrated player scores 10/12 against avg. 1800 opponents

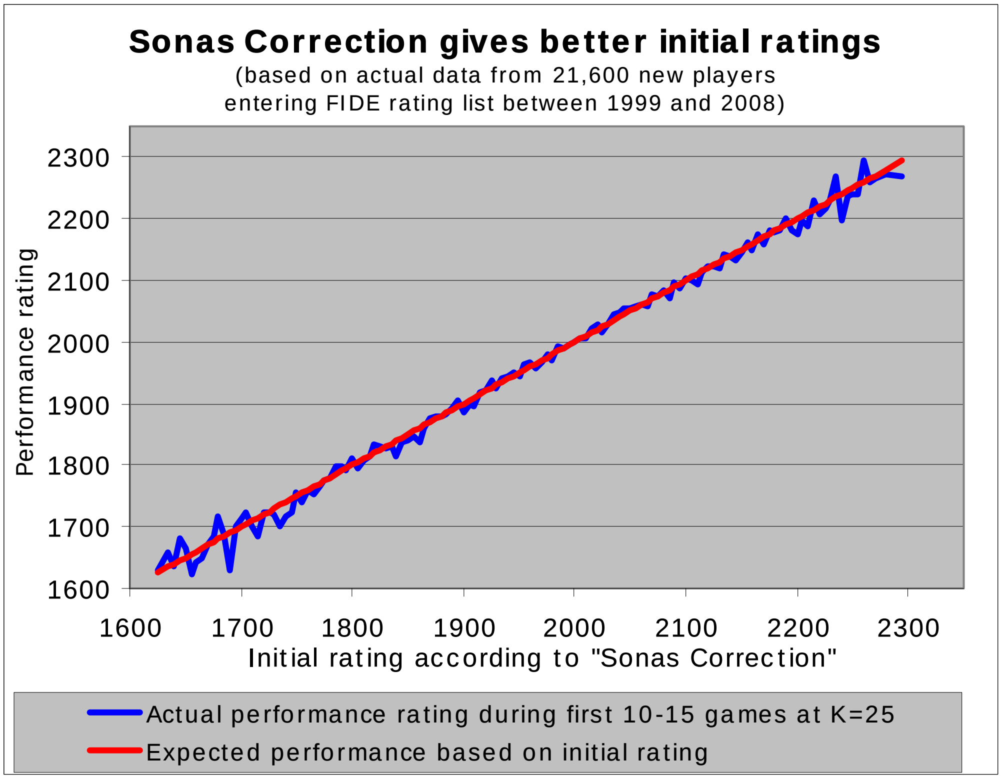
(A)
Original approach: 2/9 is 22%, which corresponds to -220 rating points. So Player A gets an initial rating of 2100-220 = **1880**
New approach: real results were 2/9, an additional 4/10 is added on, for a total of 6/19 (32%), which corresponds to -133 rating points. Opponents are also considered to be 2100-40=2060. So Player A gets an initial rating of 2060-133 = **1927**

(B)
Original approach: 4.5/9 is 50%, so Player A gets an initial rating of **1950**
New approach: real results were 4.5/9, an additional 4/10 is added on, for a total of 8.5/19 (45%), which corresponds to -36 rating points. Opponents are also considered to be 1950-40=1910. So Player A gets an initial rating of 1910-36 = **1874**

(C)
Original approach: 5/9 is greater than 50% score, actually a +1 score, leading to a +12.5 point bonus. So Player A gets an initial rating of 2300+12.5=**2313**
New approach: real results were 5/9, an additional 4/10 is added on, for a total of 9/19 (47%), which corresponds to -21 rating points. Opponents are also considered to be 2300-40=2260. So Player A gets an initial rating of 2260-21 = **2239**

(D)
Original approach: 10/12 is greater than 50% score, actually a +8 score, leading to a +100 point bonus (8x12.5=100). So Player A gets an initial rating of 1800+100=**1900**
New approach: real results were 10/12, an additional 4/10 is added on, for a total of 14/22 (64%), which corresponds to +102 rating points. Opponents are also considered to be 1800-40=1760. So Player A gets an initial rating of 1760+100=**1860**

As you might expect, this formula was specifically optimized so that the red and blue lines match up in the graph, meaning that no matter what your initial rating is, it is a very good prediction of what your performance rating will be over your next 10-15 games. Compare that last graph against the results provided by the "Sonas Correction":



The justification for adding the extra 4/10 score is that it will push players' calculated percentage scores closer to 40%, and therefore will have the effect of clustering the initial ratings more closely around the 1800-2000 range (under this formula, the frequency of players getting an initial rating between 1800 and 2000 will increase from its current value of 33% up to 42%). And the justification for subtracting 40 points from the strength of opponents is that rated opponents do not seem to play up to their normal strength when facing unrated players.