CSE702, Spring 2024: Analyzing Cognitive Tendencies From Chess Data

We will begin with metrics for cognitive skills, in cless and otherwise. The initial parts will also include a tour of the large data for the chess-based metrics, much of which is on private pages.

Elo Ratings

Elo ratings, which were devised by the Hungarian-American statistician Arpad Elo, are a measure of skill---like GPA in academics. They originated in chess but have been applied in sports, online role-playing games, even (for a short time) Tinder.

Most chess federations use a scale in which fixed numbers have fixed interpretations. Originally Elo targeted **1500** to be the median of players keen enough to join the U.S. Chess Federation. He chose **200** to be the *source standard deviation*, but the actual distributions of rated players need not conform to that. The USCF also reasoned that an interval of 200 points should define a *class* of players. The classes I remember from the 1970s, ranging from 1000 to 2000 and then higher, are:

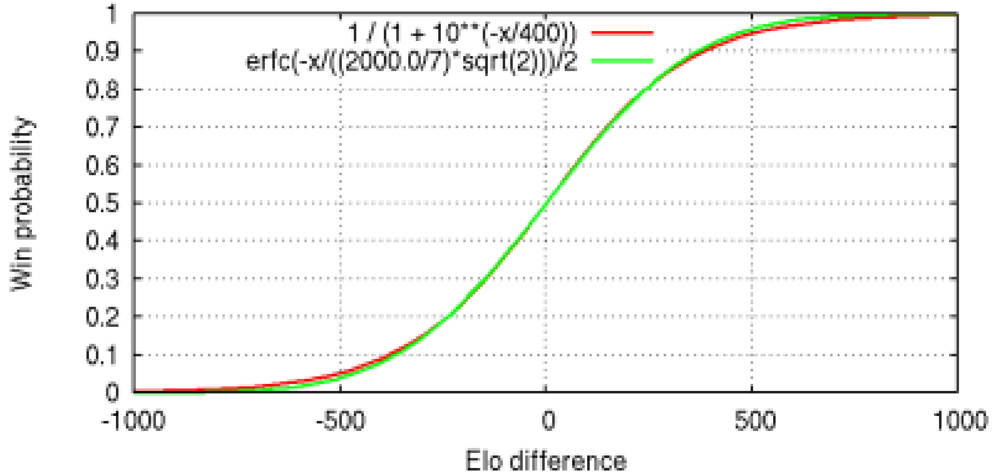
- [1000...1200): Class E
- [1200...1400): Class D
- [1400...1600): Class C
- [1600...1800): Class B
- [1800...2000): Class A
- [2000...2200): Expert
- [2200...2400): **Master**; *threshold* for FIDE Master (FM) title is 2300 (need only cross once)
- 2400+: Senior Master, highest USCF rank; threshold for FIDE International Master (IM)
- 2500+: Threshold for FIDE Grandmaster (GM) (IM and GM also need higher *norms*.)
- 2600+: Informal threshold for "Super-GM". 100th player is rated 2638; top woman is 2633.
- 2700+: "Elite". Was more than 50 players before the pandemic, now only 31. (Deflation?)

Only 13 human players have ever been rated 2800+ by FIDE, 4 now. But computer chess programs, called **engines**, reach up over **3600**. My own FIDE rating peaked at 2450 and is now 2372 (inactive).

Suppose a player P rated $R_P = 1750$ faces off against a player Q rated $R_Q = 1950$. What are the numbers supposed to mean? The system is supposed to obey the following form of "additive invariance":

The points expectation of P playing Q depends only on the difference $R_P - R_Q$.

I try to say "points expectation" rather than "win probability" because chess has drawn outcomes. So if the probability of winning is w and the probability of a draw is d then the points expectation is $w + 0.5d$. But not everyone is so careful, including the original designer of the following graphic, in which a probit curve is superposed on Elo's original logistic curve:



So a 200-point difference gives roughly 75% expectation for the stronger player Q. Exactly according to the formula it is

$$q = \frac{1}{1 + 10^{-(R_Q - R_P)/400}} = \frac{1}{1 + 10^{-0.5}} = \frac{\sqrt{10}}{1 + \sqrt{10}} = \frac{3.16227766...}{4.16227766...} \approx 0.7597...$$

which is basically 76%. Close enough to say that a "Class Unit" is an interval of 75% expectation of the stronger over the weaker player. By "additive invariance", these intervals have the same 200-point width over the whole scale from beginner to champion. So we can also say:

- Computers are 4 class units above the best human players.
- Magnus Carlsen is more than 2 class units above a typical IM. Par for the IM (like me) would be to get 1 draw in 10 games.
- But I am 2 class units above some truly brilliant 1900s-rated people who are keen at chess. Why?
- Those people are 4 class units above Class E players, who are still pretty good.
- One can reckon 600 as "adult beginner", though scholastic beginners go down below 100.
- Thus chess has 11 class units from "adult beginner" to champion. "**Depth** of Chess."
- László Mérő, in his 1990 book *Ways of Thinking* (rev. in 2004 as *Habits of Mind*) measured class units and depth for other games and sports.
- The NFL Elo Ratings by Nate Silver and others basically range from 1300 to 1700+. Silver left in 2023, but Neil Paine's ratings are up-to-date and similar.

The ratings are intended to be **predictive**. The Bills at 1704 would still be a slight underdog to the Cheifs at 1722. BTW, this does not mean that I outrank the Bills. Silver merely imitated Arpad Elo by starting the 32 NFL teams at 1500 then running a rating-update simulation according to actual won-lost records of past seasons. That the depth of the whole league is just 2 class units means that---and is reflected in that---about 1-in-20 matchups of "best vs. worst" feature an upset.

What happens when an underdog P wins? We reason that P was not quite so much an underdog, by upping the rating R_P. It doesn't necessarily go higher than R_Q or even up to R_Q, but it definitely closes some of the difference $R_Q - R_P$. The magnitude of the update is controlled by a customizable parameter conventionally called K. The update can be applied in one go over any set of games, and the resulting new rating R'_P is defined by:

$$R'_P \;=\; R_P \;+\; K \cdot (actualPts \;-\; projectedPts).$$

The projected points are added up for each game. The rating updates can themselves be applied individually to each game---and then the end result is generally *not* the same as doing the single update on the block of games. But it is usually close. The projected points depend only on the differences $R_P - R_Q$ over each opponent Q, and hence so does the amount of rating points gained or lost.

Thus Elo ratings are expressly relative. But the chess scale is intended to be absolute with regard to skill. **How steady are its mileposts?** That is an item for us to explore...

GPA-Like Individual Metrics

GPA is also predictive. It is supposed to predict real-world success. (!!??) Of most immediate concern locally is how well it predicts scores on projects, homeworks, and (especially) examinations. It is also supposed to be absolute, though the actual levels vary between academic institutions.

The main difference between GPA and Elo is that students do not compete against each other in the *zero-sum* manner of chess. Same thing with job-performance measures, and **psychometrics** quite in general. In chess we can define other non-competitive skill metrics.

- From long ago there have been puzzle-solving ratings.
- With computer programs strong enough to be regarded as giving objective absolute values of move decisions, we can define many new "objective" skill metrics.

"Objective" means raw counting, with no dependence on model training. With respect to a strong computer chess engine and span of search used as the benchmark, here are some core metrics:

- **T1**-Match (called **MMP** for Move-Match % by me): The % of playing the move listed first by the engine.
- **EV**-Match: includes a move of **e**qual **v**alue to the first move as a match. (Recommended in this paper, which called it CV for "coincidence value.") (Only for the first 5 listed moves.)
- **ACPL**: "Average Centipawn Loss"---means without scaling.
- **ASD**: Average Scaled Difference---see "When Data Serves Turkey" versus ACPL.
- **Err025**: Count of errors of 0.25 or more (not scaled).

There are others: **T3-Match** credits any of the first 3 listed moves by the engine, while **T3thr50** credits playing a top-3 move only if at most 0.50 inferior (not scaled). They are motivated by the thought that "smart cheaters" often play 2nd-best or 3rd-best moves to throw off detection via T1 or EV, provided the move is not too bad. Maybe **T3thr50** is the most reflective of chess skill overall among these metrics, but that is verging on the purpose of the "full model" to come. Let's stick with the above simple-counting metrics for now.

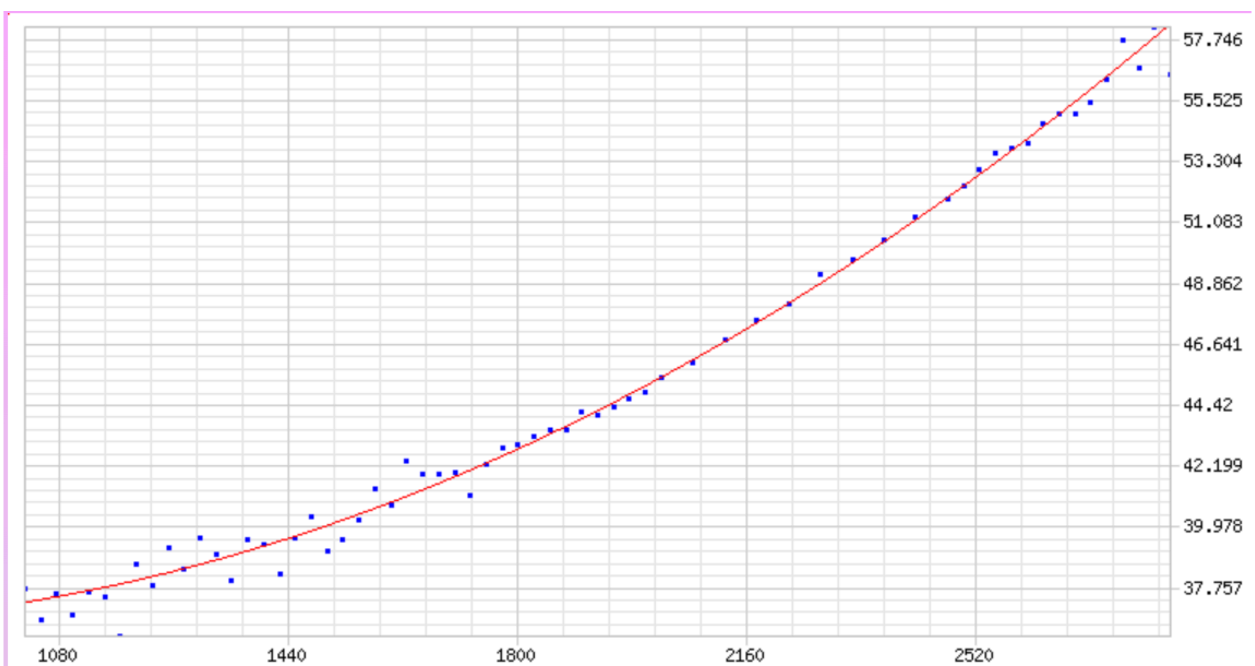
Our question is: **Should these quantities be strictly linear in the chess ratings of the players, all across the scale from neophyte to champion?** (And up to computers---up to the limits of resolution from programs themselves being the benchmark.) This presupposes that the population of players is in a good "steady state" with regard to ratings. This may fail for several reasons:

1. The update rule $R' = R + K \cdot (actual - proj.)$ may not "mix" fast enough.
2. The population is not static: people (such as myself) leave having withdrawn more points than given, especially at the high end---but there is even more turnover at the low end.
3. External events may derail the correspondence of rating to skill---war/isolation/pandemic.

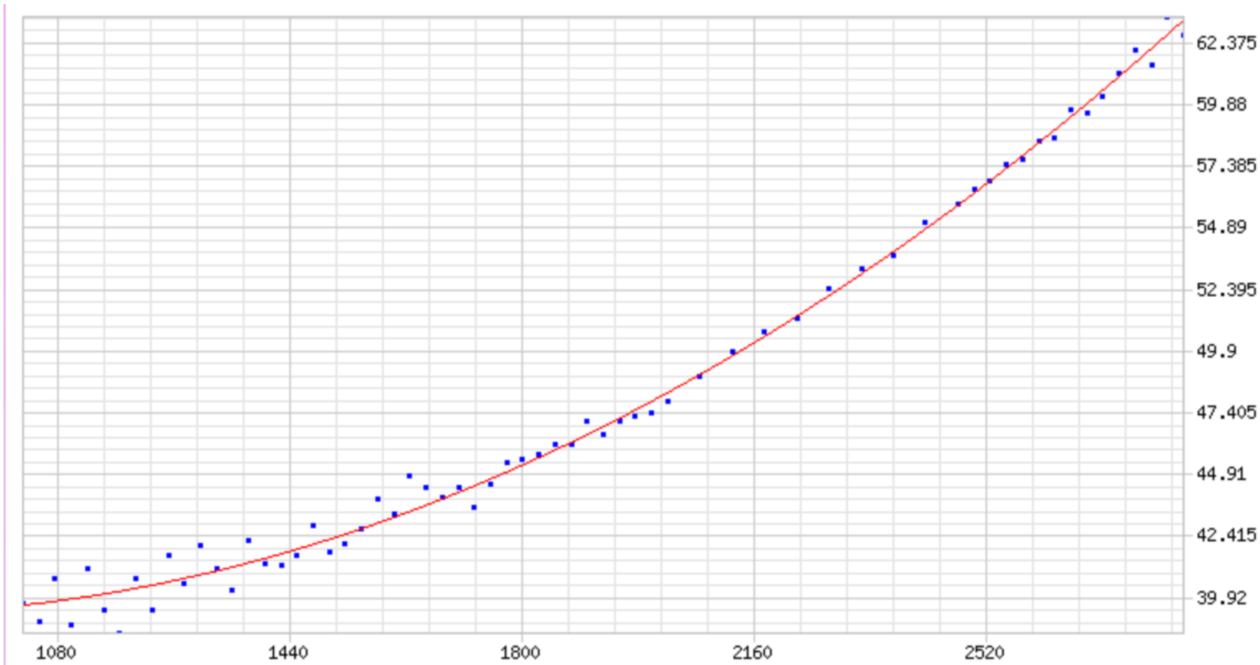
Graphs of These Metrics Versus Elo Ratings In 2010--2019

Prior to the pandemic, I was mystified by the following graphs:

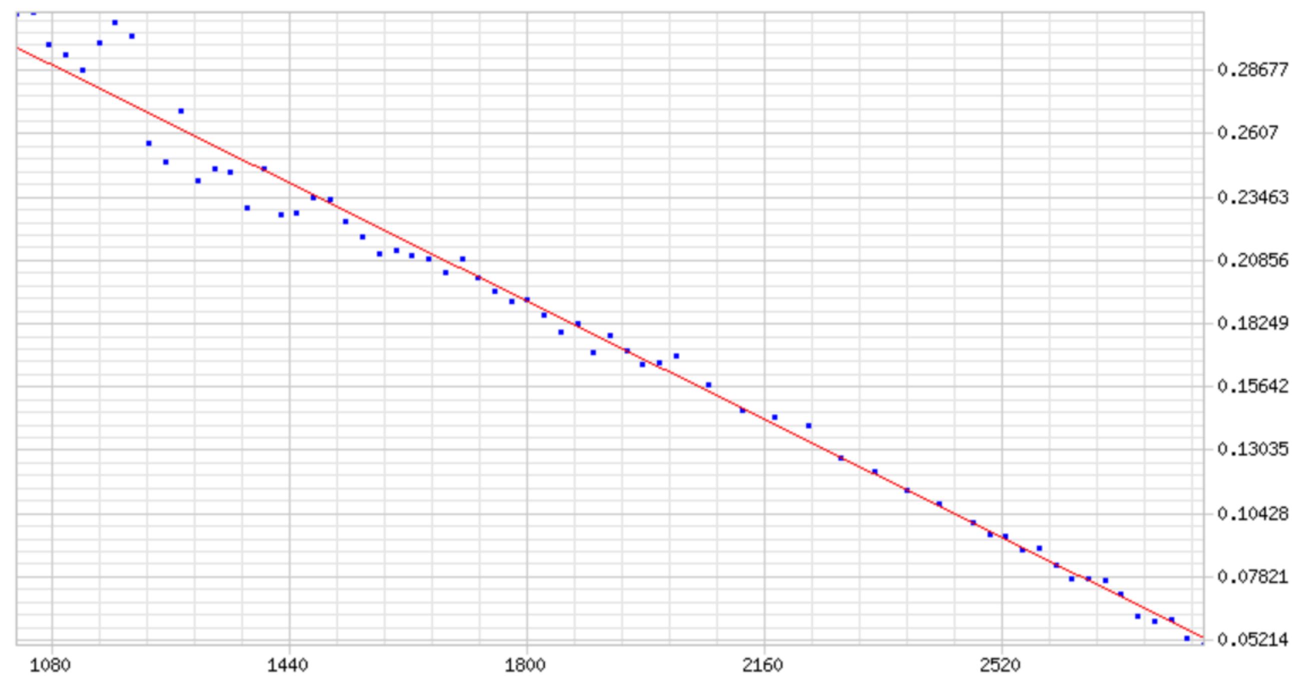
T1-Match:



EV-Match:

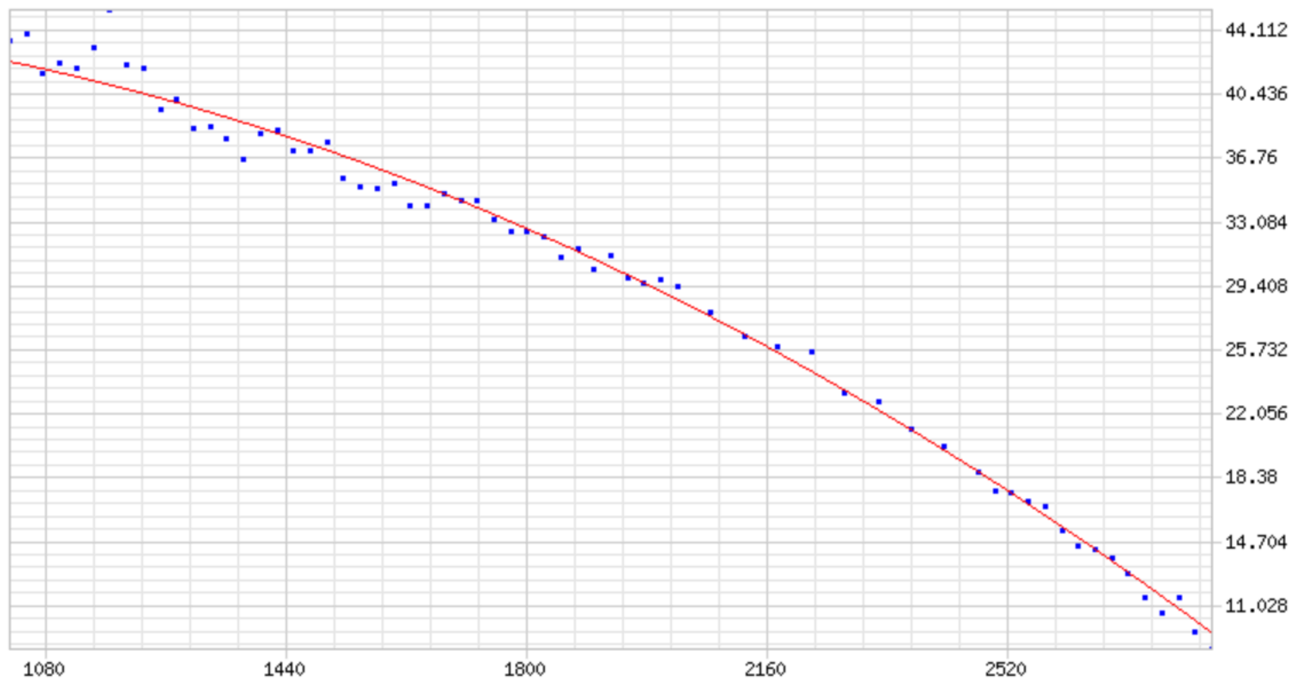


ASD: (See similar graphs with unscaled ACPL at the end.)



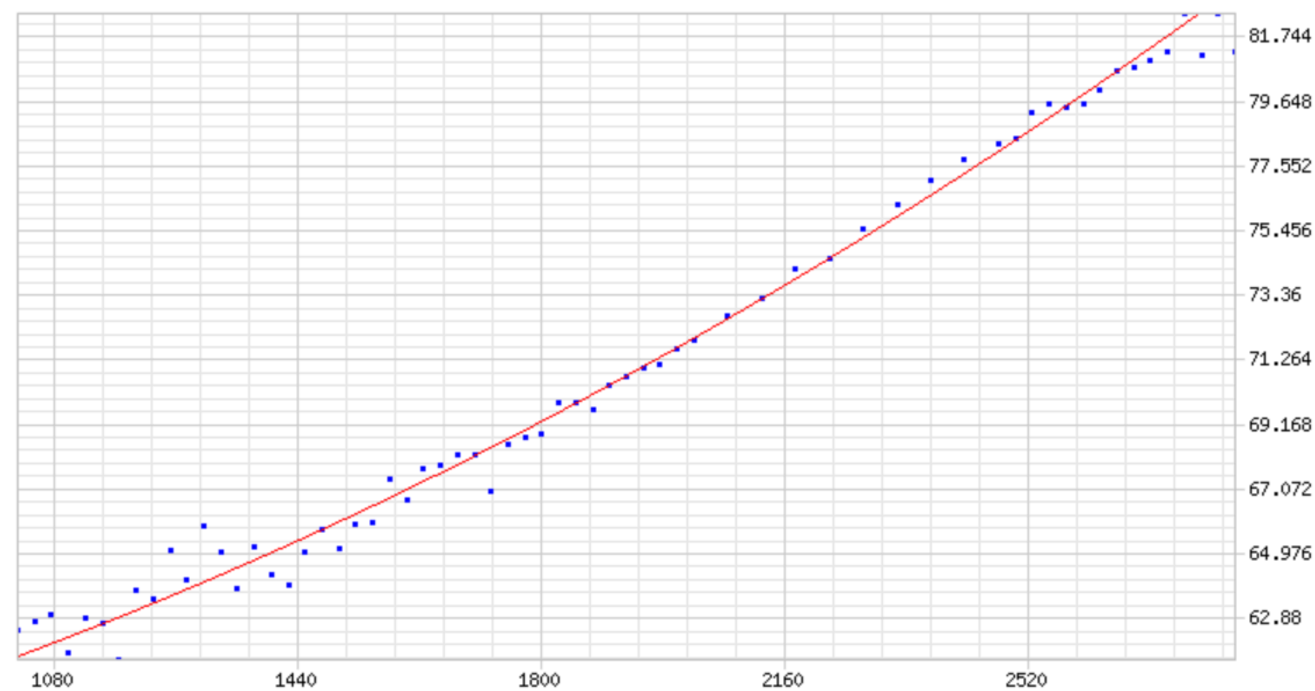
This still fits a straight line. The R^2 is 0.9861 weighted by the move sample size for each data point, 0.9863 unweighted. (This is a major reason I did not suspect the nonlinearity in T1 until 2018, well after the "Turkey" article.)

Error Count Err025:



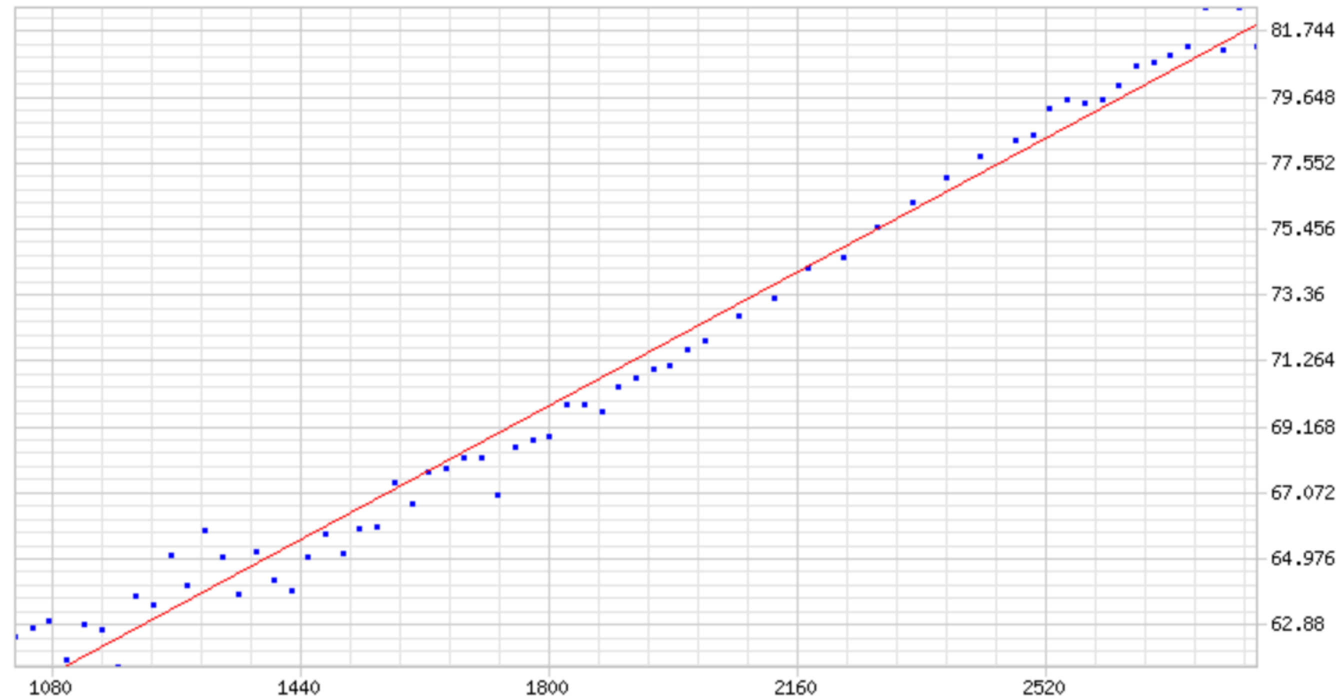
Again curved.

T3-Match:



Still curved, with $R^2 = 0.9898...$

A line, however, is not bad either---$R^2 = 0.9787...$ weighted, $0.983...$ without:



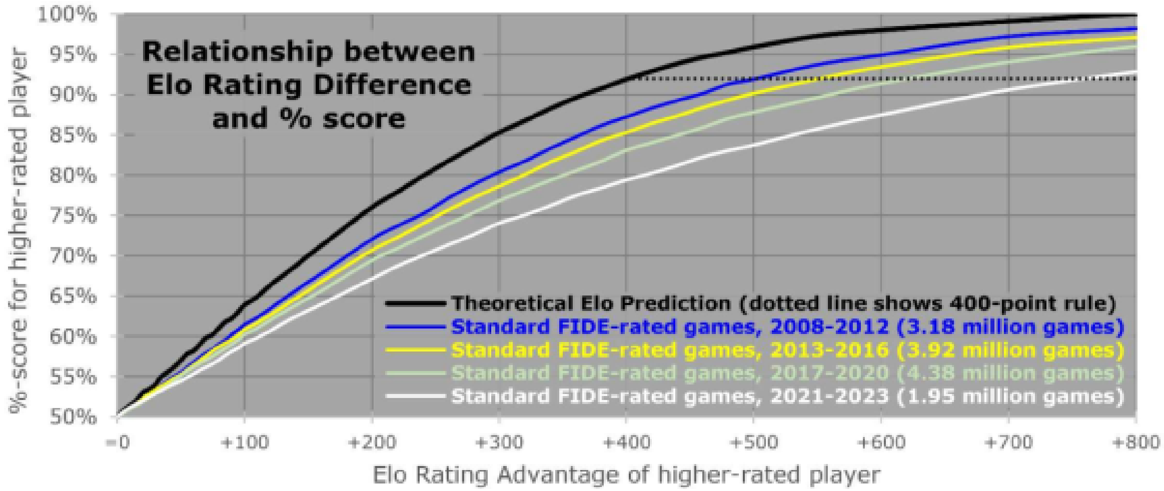
Let's just focus on T1 versus ASD, where the issue is clearest: the latter is a straight line, but the former definitely is not.

Graphs of These Metrics After the March 2024 "Sonas Correction"

The pandemic threw Elo ratings out of whack because in-person chess largely stopped and only in-person chess is FIDE-rated. Young players' minds did not stop growing, as they picked up chess knowledge and experience online. I noticed in September 2020 and quantified this in Nov.-Dec. 2020 while monitoring the FIDE World Youth Rapid Championship.

- July 2021 "Pandemic Lag" article.
- July 2023 case of Indian teen Sarayu Velpula (FIDE card). Various articles about.
- August 2024 article.

FIDE's official statistician Jeffrey Sonas explained that even long before the pandemic, the predictive accuracy of FIDE's Elo ratings had "sagged" majorly:



The discrepancy was mainly in ratings below 2000:

2008-2012	opponent / player	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+
+4.5% in 6,865 games	1200-99		+1%	+3%	+6%	+6%	+7%	+4%	+4%	+2%	+2%	=0%	=0%	=0%	=0%
+4.8% in 25,544 games	1300-99	-1%		+2%	+4%	+7%	+7%	+4%	+5%	+4%	+4%	+1%	=0%	+2%	=0%
+5.0% in 60,420 games	1400-99	-3%	-2%		+3%	+5%	+7%	+8%	+7%	+4%	+3%	+1%	+1%	+1%	=0%
+4.9% in 135,409 games	1500-99	-6%	-4%	-3%		+3%	+6%	+7%	+8%	+6%	+5%	+3%	+2%	+1%	=0%
+4.0% in 241,595 games	1600-99	-6%	-7%	-5%	-3%		+4%	+6%	+8%	+7%	+6%	+4%	+2%	+2%	+0%
+2.7% in 385,143 games	1700-99	-7%	-7%	-7%	-6%	-4%		+3%	+6%	+7%	+6%	+5%	+3%	+1%	+1%
+1.4% in 540,108 games	1800-99	-4%	-4%	-8%	-7%	-6%	-3%		+3%	+5%	+6%	+4%	+4%	+2%	+1%
+0.0% in 699,217 games	1900-99	-4%	-5%	-7%	-8%	-8%	-6%	-3%		+3%	+4%	+4%	+3%	+2%	+0%
-1.0% in 801,405 games	2000-99	-2%	-4%	-4%	-6%	-7%	-7%	-5%	-3%		+2%	+3%	+3%	+2%	+1%
-1.5% in 779,484 games	2100-99	-2%	-4%	-3%	-5%	-6%	-6%	-6%	-4%	-2%		+2%	+2%	+2%	+1%
-1.3% in 644,501 games	2200-99	=0%	-1%	-1%	-3%	-4%	-5%	-4%	-4%	-3%	-2%		+2%	+2%	+1%
-1.3% in 453,537 games	2300-99	=0%	=0%	-1%	-2%	-2%	-3%	-4%	-3%	-3%	-2%	-2%		+2%	+1%
-1.1% in 315,497 games	2400-99	=0%	-2%	-1%	-1%	-2%	-1%	-2%	-2%	-2%	-2%	-2%	-2%		+1%
-0.8% in 213,368 games	2500+	=0%	=0%	=0%	=0%	-0%	-1%	-1%	-0%	-1%	-1%	-1%	-1%	-1%	

His fix was to do a one-time shift on the interval [1000,2000] onto [1400,2000]. Thus a player rated 1000---then the FIDE minimum---instantly became 1400. A 1500-rated player became 1700; a 1900-rated player became 1940. Players rated above 2000 were not affected. This change was approved and took place on March 1, 2024.

How has it worked since then? I started with July 1, 2024, to allow ratings to adjust more, including summer events in June, and ran until December 31. Results:

T1-Match:

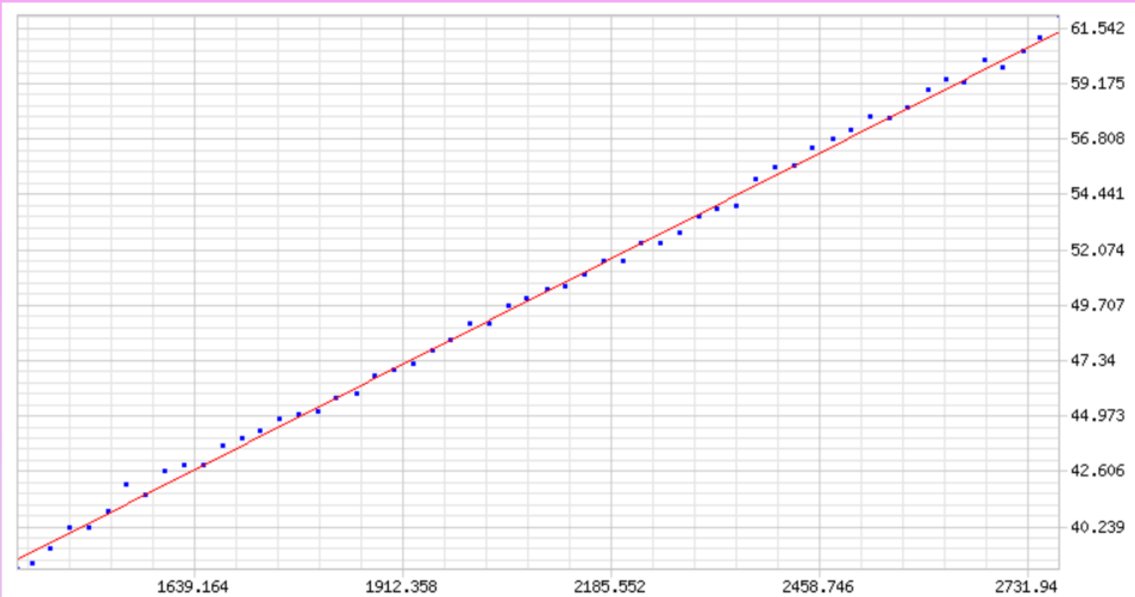
Function

$f(x) = 15.629366010177451 + 0.016505383545018244x$

R-Squared

$R^2 = 0.99798640446163$

Graph



Letter-perfect! But

ASD:

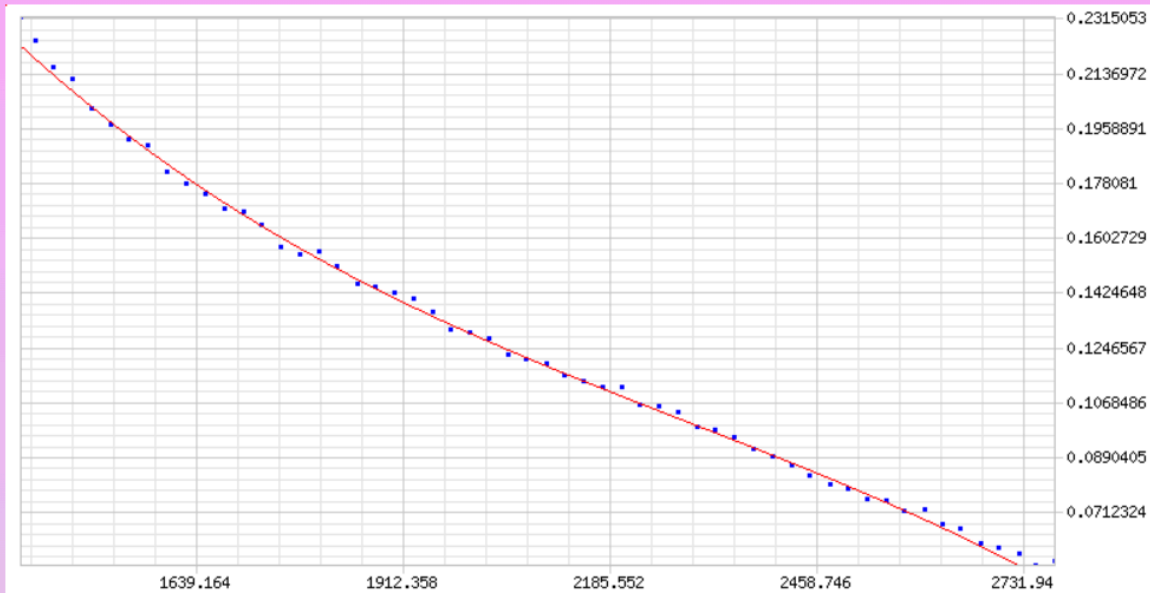
Function

$$f(x) = 0.9860953066897238 - 0.000970747480041683x + 3.828945231562e\text{-}7x^2 - 5.584447936e\text{-}11x^3$$

R-Squared

$$R^2 = 0.99740836931688$$

Graph



Still curved---below 1800. This could be an artifact of my scaling formula. In general, I am happy with the change.

Titled Tuesday Blitz using FIDE Standard Ratings:

T1-Match:
(without scrubbing)

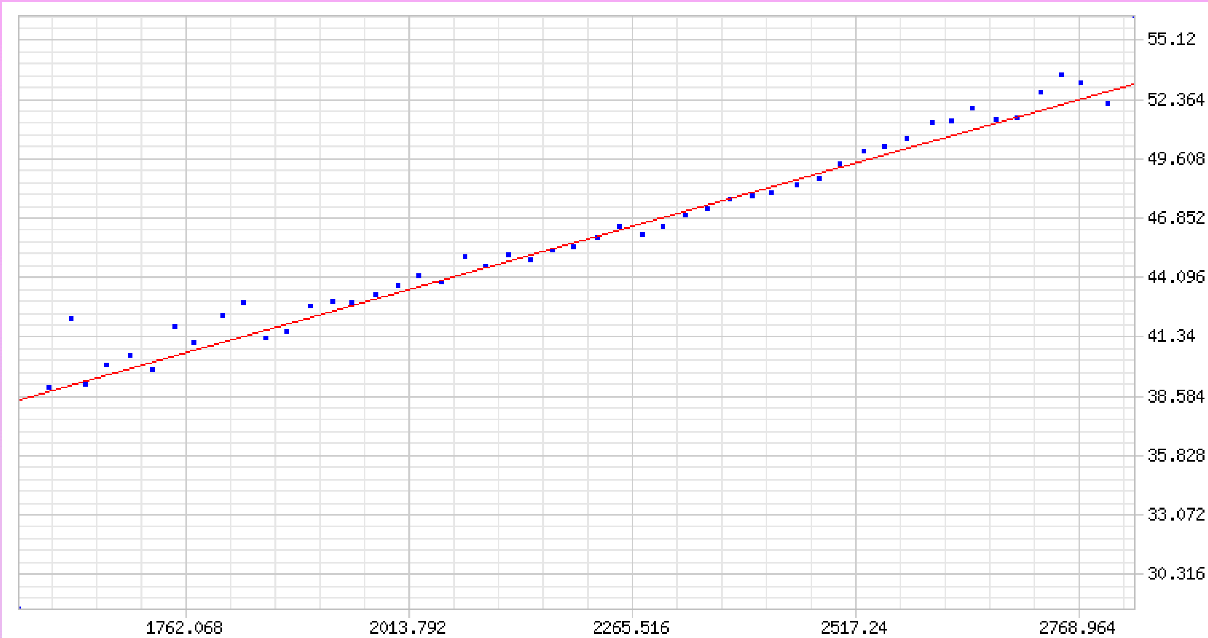
Function

$f(x) = 20.043049875967274 + 0.011670146791934807x$

R-Squared

$R^2 = 0.89408372579137$

Graph



Clear case to scrub the outlier in the lower-left corner (which is a bucket of only 108 moves) and at upper right (which is entirely Magnuis Carlsen). This yields:

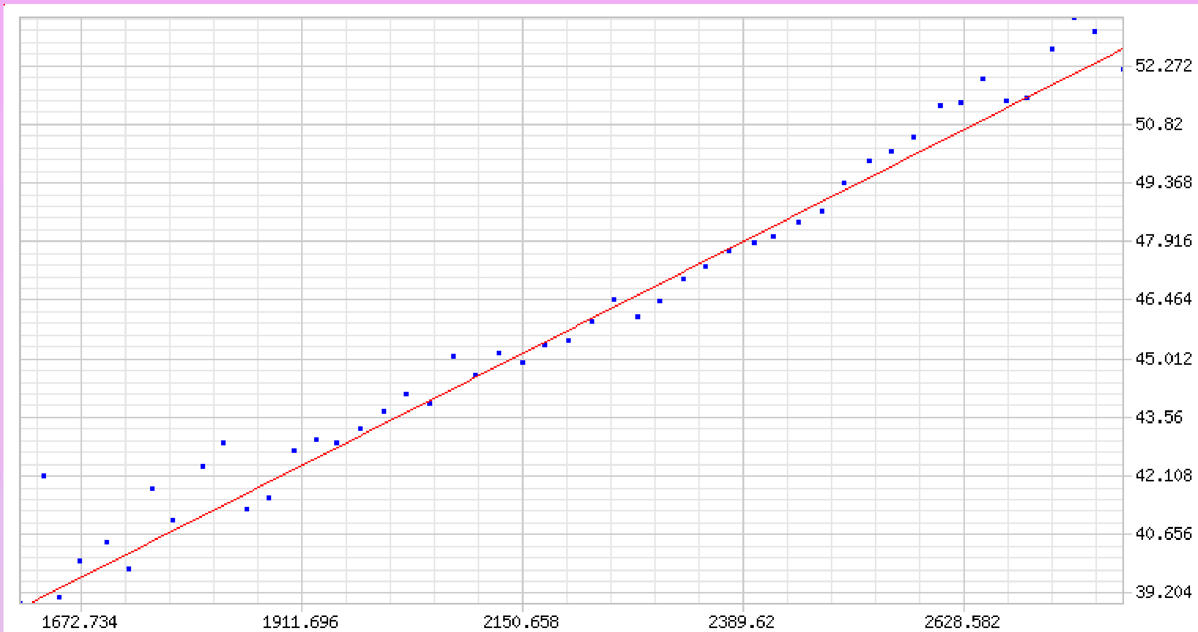
Function

$f(x) = 20.19355555557162 + 0.011601839757960737x$

R-Squared

$R^2 = 0.97024810190903$

Graph



Maybe really curved? Quadratic fit:

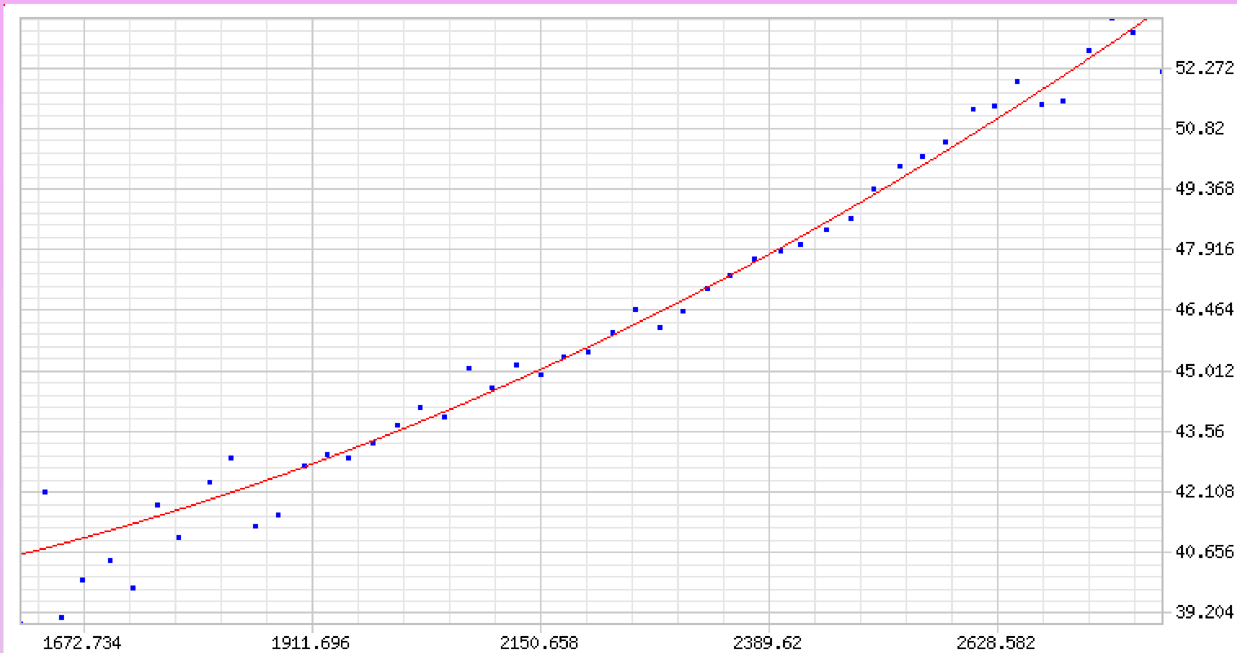
Function

$f(x) = 42.408793310623984 - 0.008053384769481562x + 0.00000431611595361724x^2$

R-Squared

$R^2 = 0.97447803619739$

Graph



The R^2 value improved only marginally. Maybe not enough justification to say curved. Plus, the data below 1900, which could be "scrubbed", actually tends more towards a line.

ASD: no scrubbing:

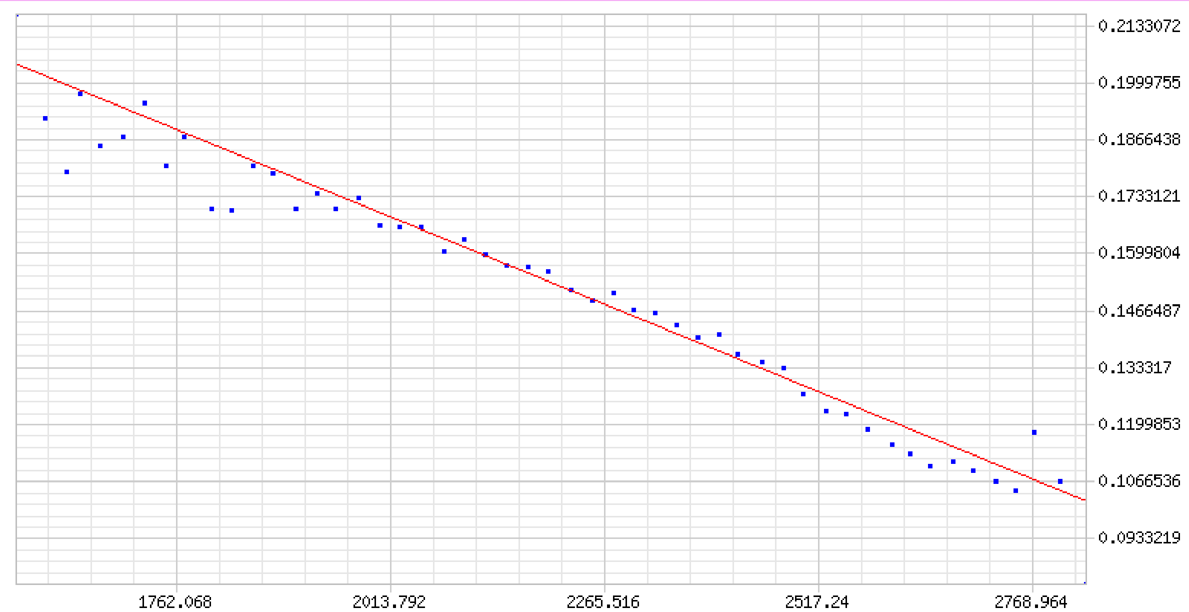
Function

$f(x) = 0.33213526753389744 - 0.000081216113384002\,58x$

R-Squared

$R^2 = 0.94935786222296$

Graph



Scrub Carlsen at lower right and small bucket at upper left:

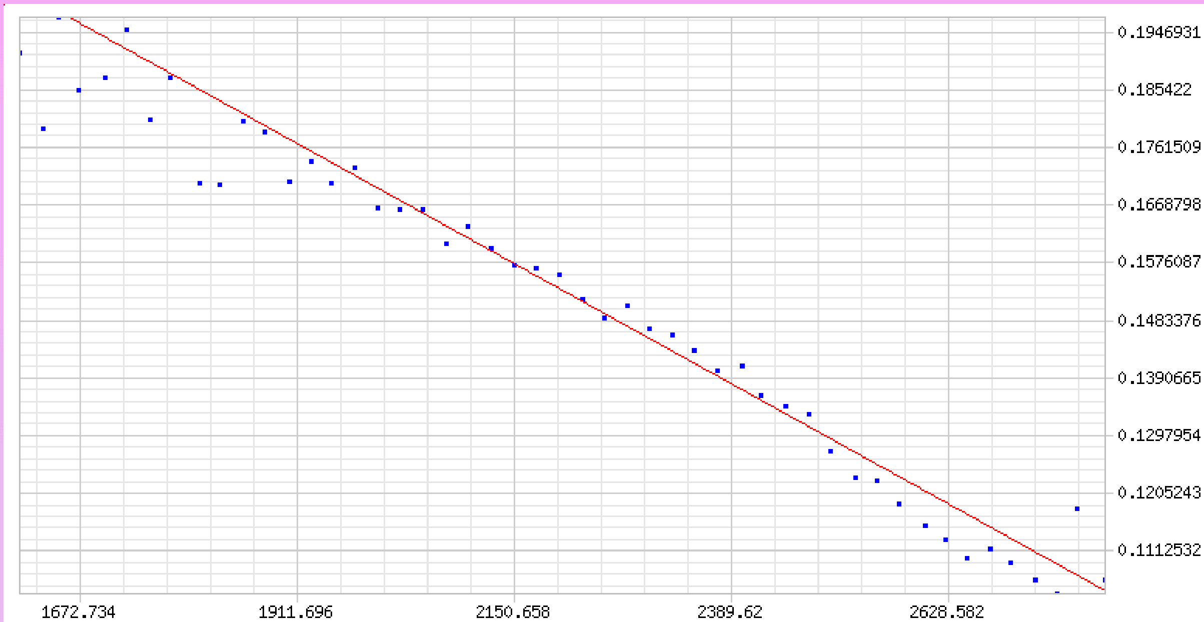
Function

f(x) = 0.331240954040705 - 0.00008080959805501638x

R-Squared

R^2 = 0.95195871358431

Graph



Looks messier. Maybe really quadratic?

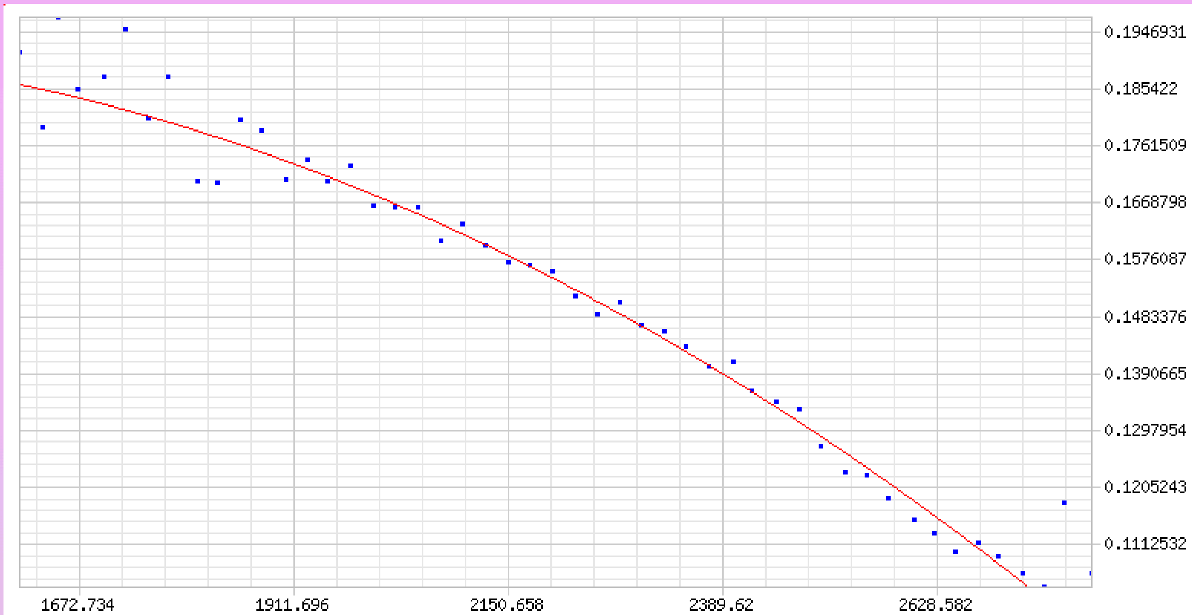
Function

$f(x) = 0.1400851458289464 + 0.0000883180237851919 3x - 3.71389513156e\text{-}8x^2$

R-Squared

$R^2 = 0.96678008015397$

Graph



A more-substantial improvement in R^2. But still messy. Scrubbing two more points on the right and below 1900 on the left:

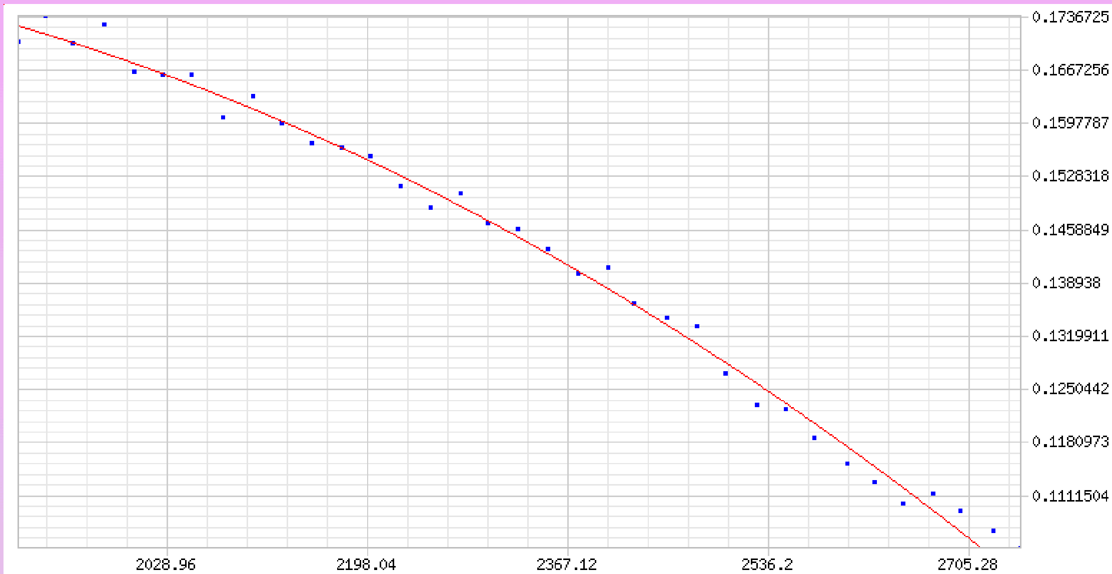
Function

$f(x) = 0.08710112356245207 + 0.00013515037298763674x - 4.743891279016e\text{-}8x^2$

R-Squared

$R^2 = 0.99171945682363$

Graph



Beautiful. But how bad is trying a straight-line fit?

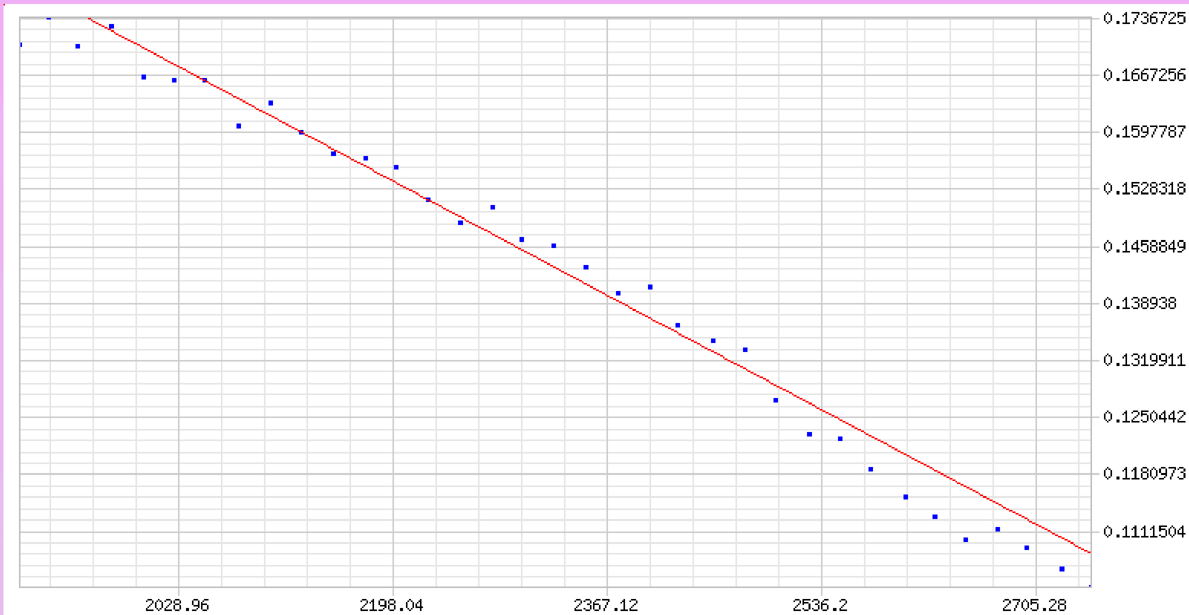
Function

f(x) = 0.33488009760335113 - 0.00008238004797429378x

R-Squared

R^2 = 0.9774949166846

Graph



How about with respect to Chess.Com's own rating system? July---Dec. 2024 with Stockfish 16 only:

T1-Match:

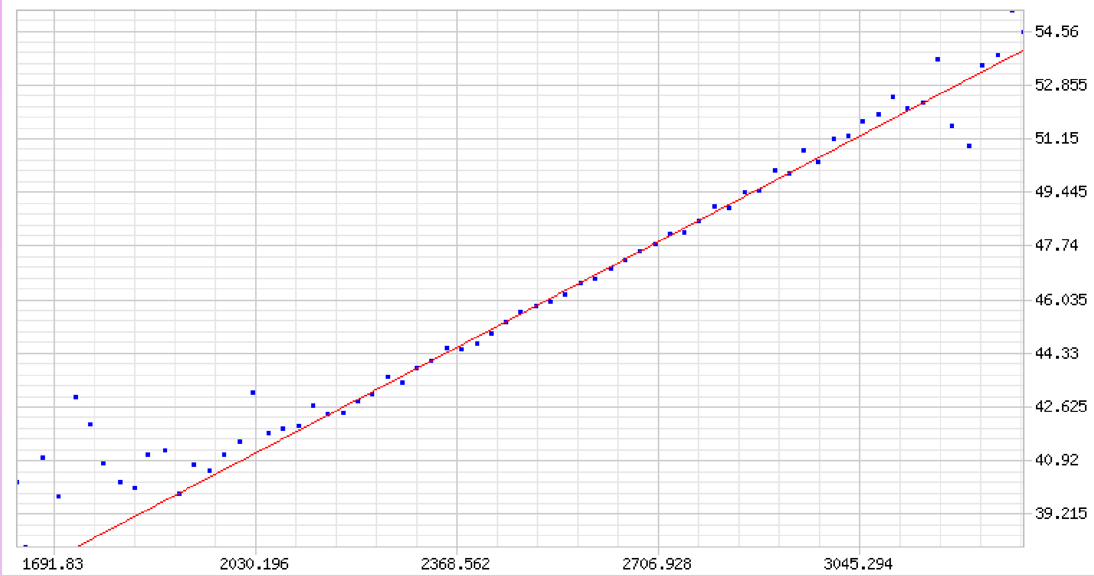
Function

$f(x) = 21.029998286995216 + 0.009922341929457797x$

R-Squared

$R^2 = 0.93425977154575$

Graph



Chess.com ratings go over 3300 and are overall inflated over 200 relative to FIDE. Since TitledTuesday has mainly upper-tier players, can scrub the small buckets for ratings below 1900:

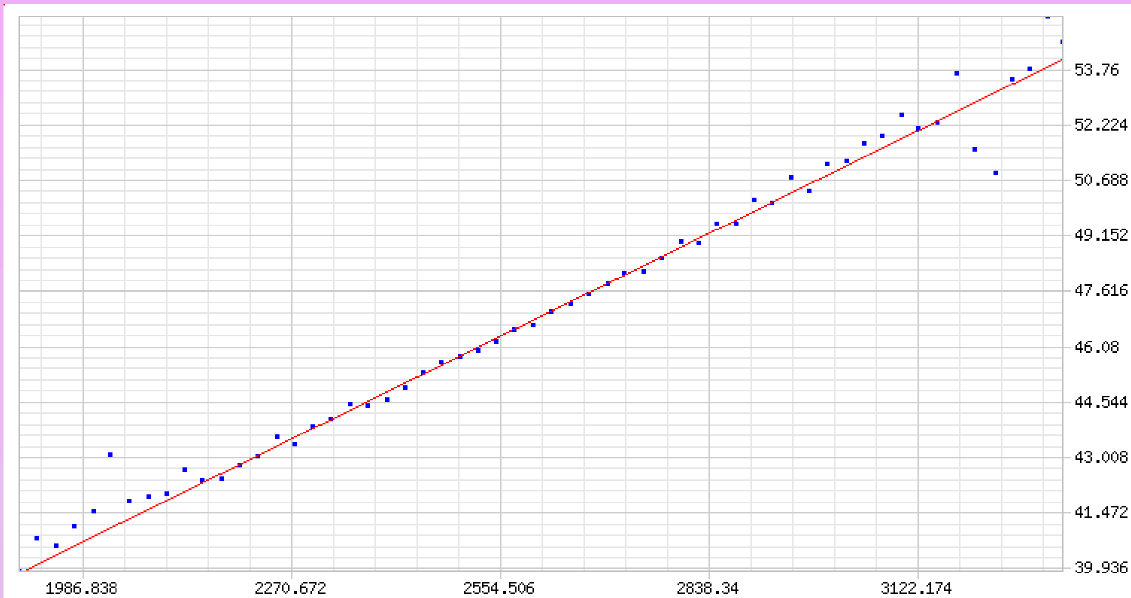
Function

$f(x) = 20.744179619963543 + 0.010030429266715631x$

R-Squared

$R^2 = 0.98223689085115$

Graph



ASD:

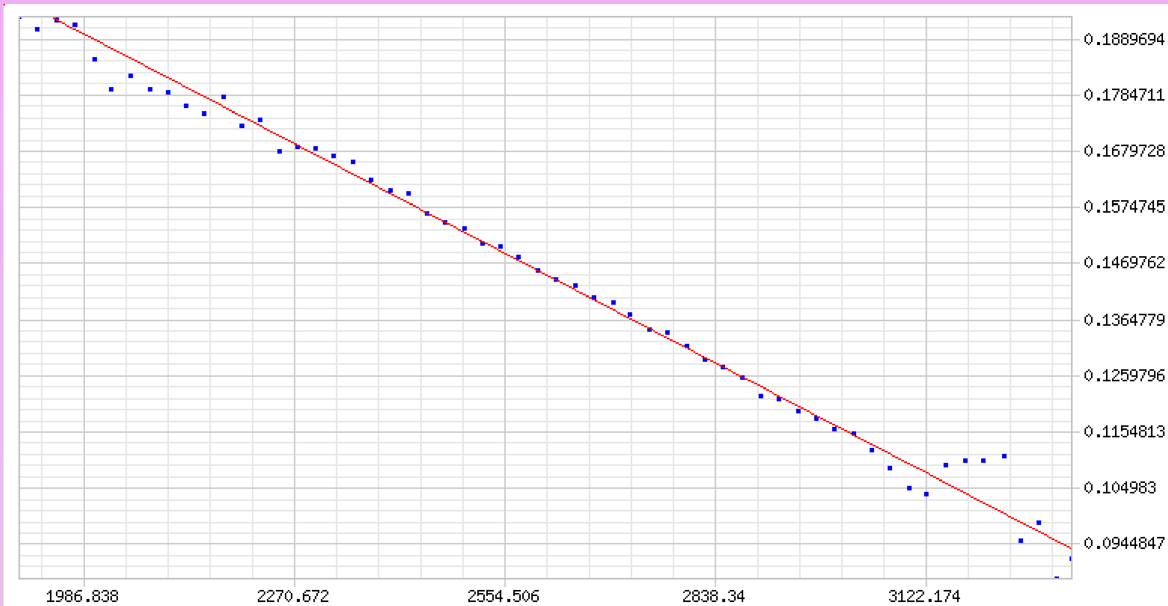
Function

f(x) = 0.3335626404346963 - 0.0000722641954127607x

R-Squared

R^2 = 0.98943308593522

Graph



Looks great too. So maybe MMP and ASD really should be linear, and there are lingering issues still with FIDE's ratings... After tonight I will see if January 2025 brings any improvement...

Tour of Sites With T1 and ASD "Screening Data"

Uniqueness of the Elo Formula?---Elementary musings based on an "Angry Statistician" post:

Suppose Player 1 has probability x of beating a generic opponent and Player 2 has probability y. Can we infer from x and y the probability p of Player 1 beating Player 2? We have some axioms:

1. $x = y \implies p = 0.5$.
2. $x = 0 \implies p = 0$ (maybe unless $y = 0$).
3. $x = 1 \implies p = 1$ (maybe unless $y = 1$).
4. $y = 0 \implies p = 1$ (maybe unless $x = 0$).
5. $y = 1 \implies p = 0$ (maybe unless $x = 1$).

It turns out we can derive a formula $p(x,y)$ with this behavior by dividing the "Player 1 odds ratio" $\frac{x}{1-x}$ by the Player 2 ratio $\frac{y}{1-y}$ to solve for the "direct confrontation odds ratio":

$$\frac{p}{1-p} = \frac{x(1-y)}{y(1-x)}$$

You can think of the odds ratio as the amount of money you need to bet to win $1 when the payoff reflects the probability p. For instance, if $p = 0.75$ then the odds ratio is 3. If you bet $1 and win the fair payoff is $0.33... So you need to bet $3 to win $1 at this rate. Solving this for p gives $py(1-x) = x(1-y) - px(1-y)$, so $p(y - yx + x - xy) = x(1-y)$, so

$$p(x,y) = \frac{x(1-y)}{x+y-2xy}.$$

[I verified in class that this satisfies the five axioms. See interesting question in notes at the end about the extent to which this formula may be unique according to the five axioms.]

We can actually derive this formula in a more elementary way that also takes into account the idea of an incremental struggle.

Consider the following possibilities for (1) a bowler in cricket or pitcher in baseball, versus (2) a ~~batsman~~ batter:

- Bowler/pitcher makes a good delivery: probability p_1.
- Bowler has poor length/pitcher "hangs" a curveball: $q_1 = (1 - p_1)$.
- Batter has good stroke, makes solid contact: p_2.
- Batter nicks or misses ball: q_2.

For sake of argument, we suppose that if both the delivery and the batter's stroke are good, the result is a dot-ball in cricket, or a foul ball in baseball, and the confrontation goes on. This is like both players making a good move at one game turn at chess. Or if the delivery and stroke are both bad, a mistimed hit (for no runs) or another foul ball may result. We get a *result* only when:

- Batter punishes a poor delivery: boundary or home run, probability $p_1 q_2$.
- Batter fails on a good delivery: wicket or strikeout, probability $p_2 q_1$.

The probability of the batter succeeding therefore is

$$\frac{p_1 q_2}{p_1 q_2 \ + \ p_2 q_1} \ = \ \frac{p_1(1-p_2)}{p_1(1-p_2) \ + \ p_2(1-p_1)} \ = \frac{p_1(1-p_2)}{p_1 \ + \ p_2 \ - \ 2\,p_1 p_2} \ .$$

This is the same formula as before with p_1 in place of x and p_2 in place of y.

Now we note a further twist. Divide both the numerator and denominator of the leftmost form of the equation by $p_1 q_2$. This gives the overall win probability of player 1 as:

$$\frac{1}{1 \ + \ \frac{p_2 q_1}{p_1 q_2}} \ = \ \frac{1}{1 \ + \ \left(\frac{p_2}{1-p_2}\right) / \left(\frac{p_1}{1-p_1}\right)}.$$

Now we have a ratio of two odds ratio feractions nestled inside another fraction. It looks weird, but now let's think more about the nature of an odds ratio $\frac{x}{1-x}$ as a mathematical function. It is always nonnegative and increases from zero to infinity as x goes from 0 to 1. This is the same range behavior as the exponential function e^M where M goes from $-\infty$ to $+\infty$, i.e., as a function of the whole real number line. In fact, the correspondence is exactly $M = \ln\left(\frac{x}{1-x}\right)$ which is the **logit** function, but let's not even think of that. Let's think of M abstractly as a measure of "mojo". A person who is more likely to lose than win ($x < 0.5$) has "negative mojo." An omnipotent player has infinite mojo, while a hopeless player has negative infinity mojo. If we substitute the "mojo" representations using M_1 and M_2 in place of the odds ratios for p_1 and p_2, we get:

$$\frac{1}{1 \ + \ e^{M_2} / e^{M_1}} \ = \ \frac{1}{1 \ + \ e^{(M_2 - M_1)}}.$$

The philosophical magic is this: We have converted the win probability of player 1 from a function of two variables representing the players separately into a function of *only one variable*: the "difference in mojo" between the players. This also means that the relation of winning probability to (difference in) "mojo" is the same across the scale.

(Note, incidentally, that this win probability is not meant to be the same as the "p_1" (or "x") we started with. The first time we derived the formula, x was the probability of winning against a "generic" opponent (or an average win rate over unspecified opponents), and y likewise for player 2 against general opposition; what we get is the probability p for player 1 against player 2 specifically. The second time, p_1 was a probability of personal success in isolation, which could involve skill factors apart from the quality of player 2's actions. And also by the way, we haven't yet said we are talking about *chess* or any other two-person strategy game. That chess has draws can be accommodated by the theory---we count "points expectation" instead of "win probability.")

When $M_2 > M_1$ the fraction is < 0.5, so player 1 is favored to win only when $M_1 > M_2$. What difference gives 75% win probability? Since $0.75 = \frac{1}{1+1/3}$ the answer is

$$M_1 - M_2 = \ln(3) = 1.0986...$$

Here is where I suspect that Arpad Elo, the "Martian" who converted the notion of "mojo" into a statistically regulated rating system, indulged a little bit of "numerical voodoo" to make things look cleaner for the indigenous population he landed among. Since $10^x = e^{x \ln 10}$, we can change the base to be 10 (or any other number, but the humanoids have 10 fingers). Since we haven't specified what units "mojo" comes in, let us rewrite the player 1 success formula as

$$\frac{1}{1 + 10^{(M_2 - M_1)}}.$$

Now the answer we want is $M_1 - M_2 = \ln(3)/\ln(10) = 1.0986... / 2.302585... = 0.47712...$ Hmmm...this is almost $1/2$. What happens if we plug in $M_2 - M_1 = -1/2$? We get

$$p = \frac{1}{1 + 1/\sqrt{10}} = \frac{1}{1 + 0.3162...} = \frac{1}{1.3162...} \sim= 0.7597...$$

Close enough to call this "75%"? This is so tempting, because if we want a nice round number D to mean the difference that gives "75%" probability, then our scaling factor can just be $2D$ in the denominator of the exponent, another nice round number. The US Chess Federation had already decided to call 200 points the width of a "class" under a rougher rating system devised by Kenneth Harkness in 1950, so Elo made $D = 200$ and the rating formula thus became the form it has today:

$$p = \frac{1}{1 + 10^{-(R_1 - R_2)/400}}.$$

László Mérő---who does not count as a "Martian" because he was born after WWII and stayed in Hungary---seized on the 75% advantage as a universal yardstick---a "Class Unit" of skill in any human endeavor. Being off by 0.97 percentage points may not seem a big deal, but consider this for humor: Elo's fudge is the same as considering $\sqrt{10}$ to equal 3. The Hebrew Bible has passages that seem to equate $\pi = 3$. Well, $\pi = 3.14159...$, which is less of a stretch than $3.162...$ Thus *Elo* had greater chutzpah than *Elohim*.

The formula does make 200 into the "source" standard deviation of a player's rating. If we assume that all players are equally variable in their level of "mojo" at any given time, then the standard deviation of the difference $R_1 - R_2$ becomes $200\sqrt{2}$. Elo indulged two other fudges that help everything offset well enough, the first of which most data scientists allow generally:

- The slight unevenness between a logistic curve and the "probit curve", meaning the cumulant of the normal distribution, even after the famous "1.7" scaling factor is applied.
- The approximation $\sqrt{2} = 1.41421... \approx \frac{10}{7} = 1.42857...$ That at least took rather less "chutzpah"! Thus he represented $200\sqrt{2}$ as $2000/7$.

See also Nate Solon's article "[How Elo Ratings Actually Work.]" That Elo didn't care about super-fine precision is witnessed by his famous summary of the whole shebang:

"The process of rating players can be compared to the measurement of the position of a cork bobbing up and down on the surface of agitated water with a yard stick tied to a rope which is swaying in the wind." ([quote source]).

But this is in how his system is *applied*. Speaking as a mathematical Platonist, I find the logistic formula to be **salient**---and thus "divinely ordained" as a matter of *theory*. This extends to my belief that quantities that are strong "telltales" of a player's "mojo" should be linear in it across the entire scale, full-stop.

Footnote:

An interesting further question is whether this formula is unique for **any ratio of** two possibly-infinite power series in x and y. Note that power series in just x alone encompass exponentiation and logarithms and all trig functions. So a two-dimensional power series in both x and y, and a ratio of the same, with arbitrary real coefficients, is quite a general mathematical function. To get a start on this idea, write the ratio in general form as

$$\frac{\sum_{i,j\geq 0} a_{ij}x^i y^j}{\sum_{i,j\geq 0} b_{ij}x^i y^j} = \frac{a_{00} + \sum_{i\geq 1} a_{i0}x^i + \sum_{j\geq 1} a_{0j}y^j + \sum_{i,j\geq 1} a_{ij}x^i y^j}{b_{00} + \sum_{i\geq 1} b_{i0}x^i + \sum_{j\geq 1} b_{0j}y^j + \sum_{i,j\geq 1} b_{ij}x^i y^j}.$$

Then Axiom 2 says that when $x = 0$, the entire numerator must vanish whatever y is (except that the case $y = 0$ is allowed to be indeterminate). Therefore all the coefficients $a_0 j$ with $j \geq 1$ must be identically zero, else y could make it vary. And we must have the constant term $a_{00} = 0$ too. Once you whittle down the terms this way with axioms 2--5, axiom 1 will step in to say that for each n, the sum of b_{ij} over $i + j = n$ must be exactly twice the sum of a_{ij} over $i + j = n$. Maybe it might follow that those sums must be identically zero for $n \geq 3$. Well, you could also suppose the terms with $i + j \geq 3$ are absent to begin with---i.e., that $p(x, y)$ is a ratio of quadratic polynomials. Then must the above formula be the only possibility?