Probabilities of Streaks in Online Chess

by Jeffrey S. Rosenthal, University of Toronto, www.probability.ca

(Version of September 22, 2024.)

Introduction:

Online chess has become extremely popular, especially on the web site Chess.com [1]. Such games have significant potential for cheating, since chess-playing computer programs are now much better than humans and can be easily consulted (either manually or automatically) during online matches to achieve superior play. Chess.com actively monitors and attempts to catch cheaters [2], and has recently published a detailed report of such matters [3]. Nevertheless, allegations of cheating continue, and are very controversial [4].

Some recent allegations of cheating concern long streaks of games which were all (or almost all) won by a specific player. In particular, concerns about cheating were raised [4, 5] by former world champion V. Kramnik about a recent streak of 46 games played by top-level player Hikaru Nakamura (player name: Hikaru), of which he won 45 and tied one.

I was contacted by Chess.com CEO Erik Allebest (because of my Wired interview [6]), and asked to perform an independent statistical analysis of such winning streaks. To facilitate this, I was supplied [7] with data showing all results on Chess.com of seven different top-level players, including Hikaru. I then conducted an independent statistical examination of evidence of unusual or surprising streaks in Hikaru's Chess.com game record, which I now describe. I first examined the nature of chess ratings, expected scores, win and tie probabilities, and game correlations, to establish a simple, direct model for the probabilities of various online chess outcomes.

When my original report [8] was publicized on Chess.com [9], Kramnik posted a lengthy response video and comments [10] with numerous criticisms, to which I responded in an addendum [11]; some of those issues are incorporated herein. (Perhaps unsurprisingly, Hikaru's own video response [12] was much more positive.) These same win streaks have also been examined in other contexts including a blog post [13], a Chess.com response [14], and a Bayesian perspective [15], which each reached similar conclusions through different approaches.

Of course, the existence of unlikely streaks is distinct from the issue of cheating. Indeed, a player who cheats in a consistent, regular manner might obtain a chess record which is indistinguishable from a stronger (but honest) player. Conversely, a player might perform much better over a short period due to higher concentration or motivation or preparation, even without any cheating. So, this examination should be viewed as merely investigating the existence of unlikely streaks, not the broader issue of cheating in online chess.

## Chess Ratings and Expected Scores:

One way to assess probabilities of chess outcomes is through chess ratings. Every player on Chess.com is given a chess rating, which is updated after each game. (Formally, these are "Chess.com blitz ratings"; many players also have blitz and classical ratings from the international chess federation FIDE, but we do not consider those here.)

In principle, these ratings should specify the expected score (i.e., average outcome) in each game, where the score is 1 for a win, 1/2 for a draw, or 0 for a loss. (The ratings do not, however, specify what fraction of the score should arise from wins versus draws; see next section.) Specifically, if White has rating A, and Black has rating B, so the rating difference is $D = A - B$, then one possible formula for White's expected score S is the following simple Elo logistic formula:

$$\frac{1}{1 + 10^{-D/400}} = \frac{1}{1 + 10^{-(A-B)/400}}.$$

To test the validity of this formula, we combined all games in the seven data files, and "binned" them by rating difference into bin ranges $\ldots, (-14, -5), (-4, 5), (6, 15), (16, 25), \ldots$. Then, for each bin, we computed the average score by White among all games within that bin. We compared that to the curve specified above (in red). The results were as follows:

Average Scores versus Elo Expected Scores

Rating Difference (White – Black), plus or minus 5
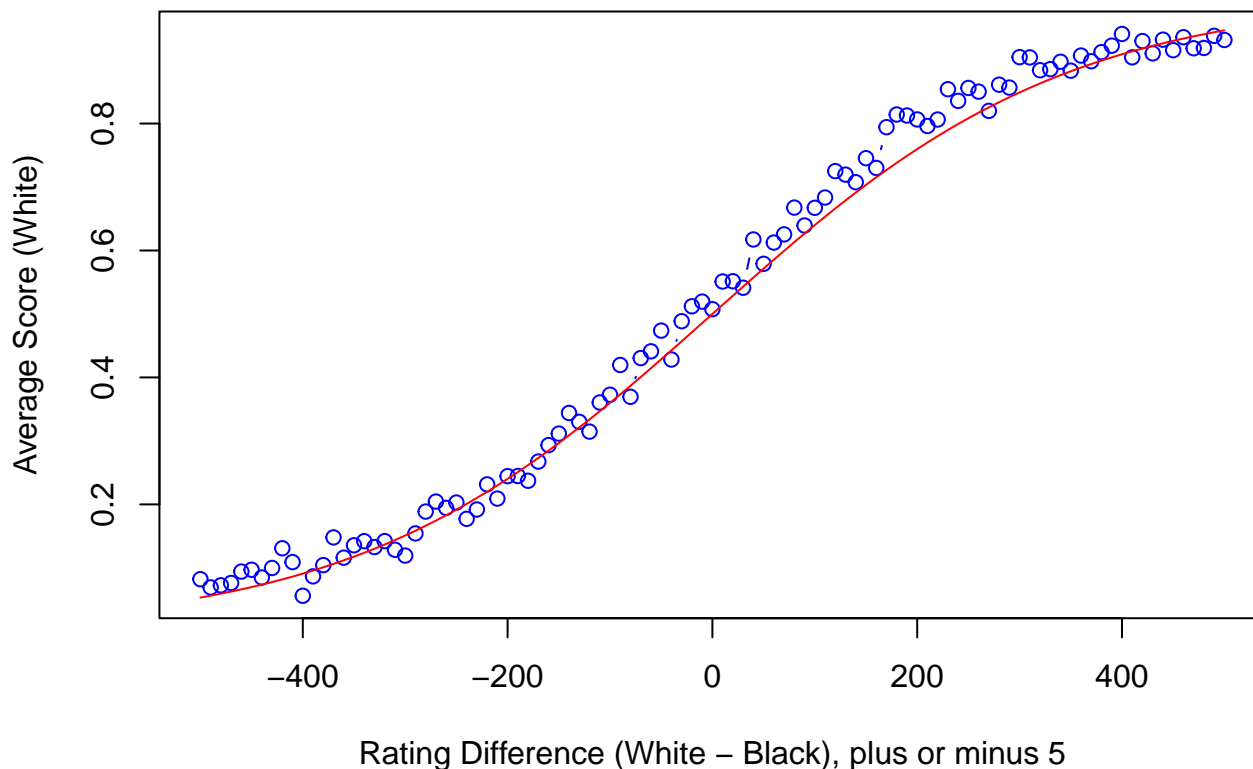
(These fit curves are presented from White's perspective. But in every game, Black's score equals one minus White's score, with rating difference the negative. So, presenting those graphs from Black's perspective would simply involve replacing X by $-X$, and Y by $1 - Y$, i.e. rotating the graph by 180 degrees, leading to exactly the same fit.)

This graph indicates that the average scores do approximately mimic the Elo expected scores, though with slight excess for rating differences between 100 and 300. Furthermore, they do not take into account the (small) advantage of playing White (i.e., moving first), even though the average score for White in the data is about 0.52 or slightly higher than 0.50. Hence, we tweaked the formula to arrive at the slightly adjusted expected score

$$S \ = \ \frac{1}{1 + 10^{-(D+14)/390}} \ = \ \frac{1}{1 + 10^{-(A-B+14)/390}} \tag{$*$}$$

(where the +14 represents White's advantage), which fits the data even better:

Average Scores versus Adjusted Expected Scores

Rating Difference (White – Black), plus or minus 5
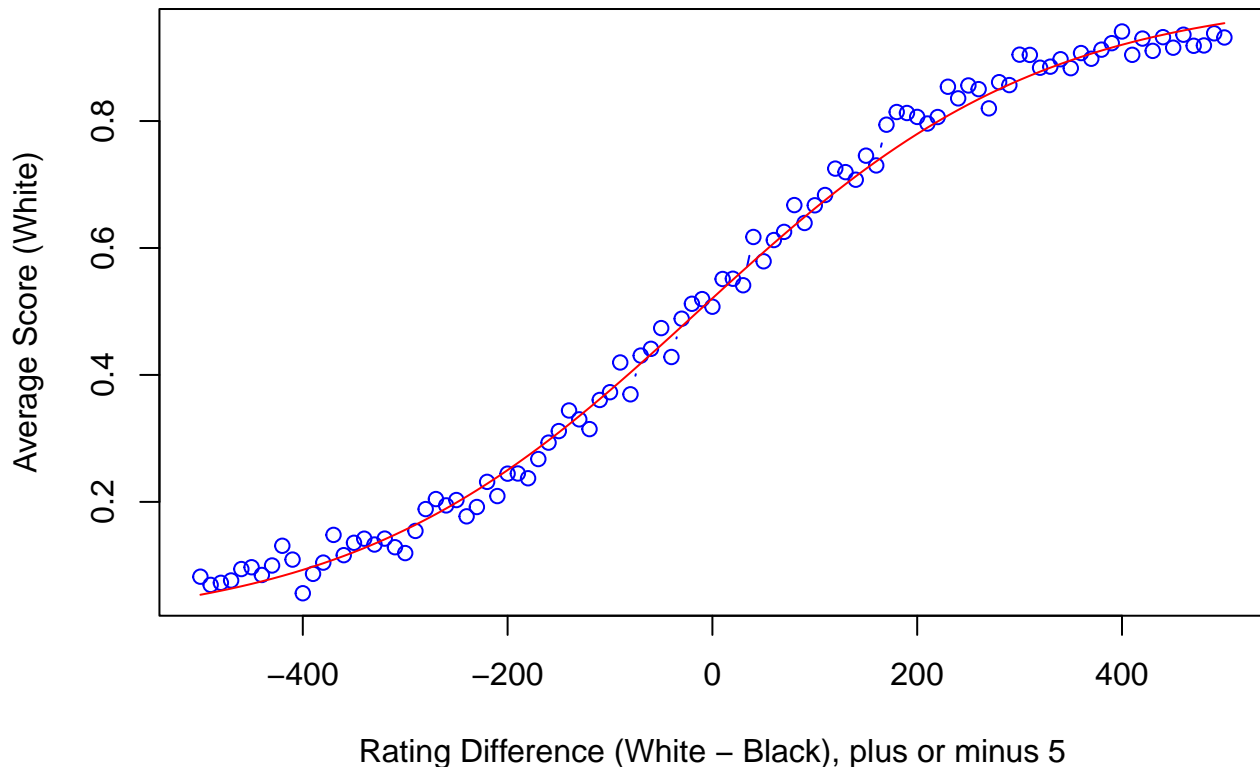
This fit is accurate enough to use to estimate probabilities. So, in our analysis below, we assume the formula ($*$) for White's expected score. (It is also possible to use the more sophisticated Glicko formulas [16], but they require knowing each player's "ratings deviation" (RD), which were not provided to me. In any case, the good fit above suggest that probability

3

estimates using (∗) are quite accurate, and any other formula which also fits the data well should give similar results.)

Draw (Tie) Probabilities:

To evaluate the likelihood of long streaks of wins and draws, it is necessary to consider not just the expected score S, but more specifically the probability W of a win and probability T of a draw (tie). Since wins give a score of 1 while ties give a score of $\frac{1}{2}$, we must have

$$S = W + \frac{1}{2}T.$$

In traditional chess tournaments with over-the-board games lasting many hours, ties are quite common. However, in online blitz chess they are less so; just 9.1% of the games in the dataset resulted in draws. Indeed, binning the data again as above, we observe that with rating difference D, the simple exponential downward-quadratic function

$$T = (1/8)\, e^{-(D/400)^2} \tag{∗∗}$$

gives a reasonably good approximation to the probability of a tie:

Probability of a Draw (Tie)

4
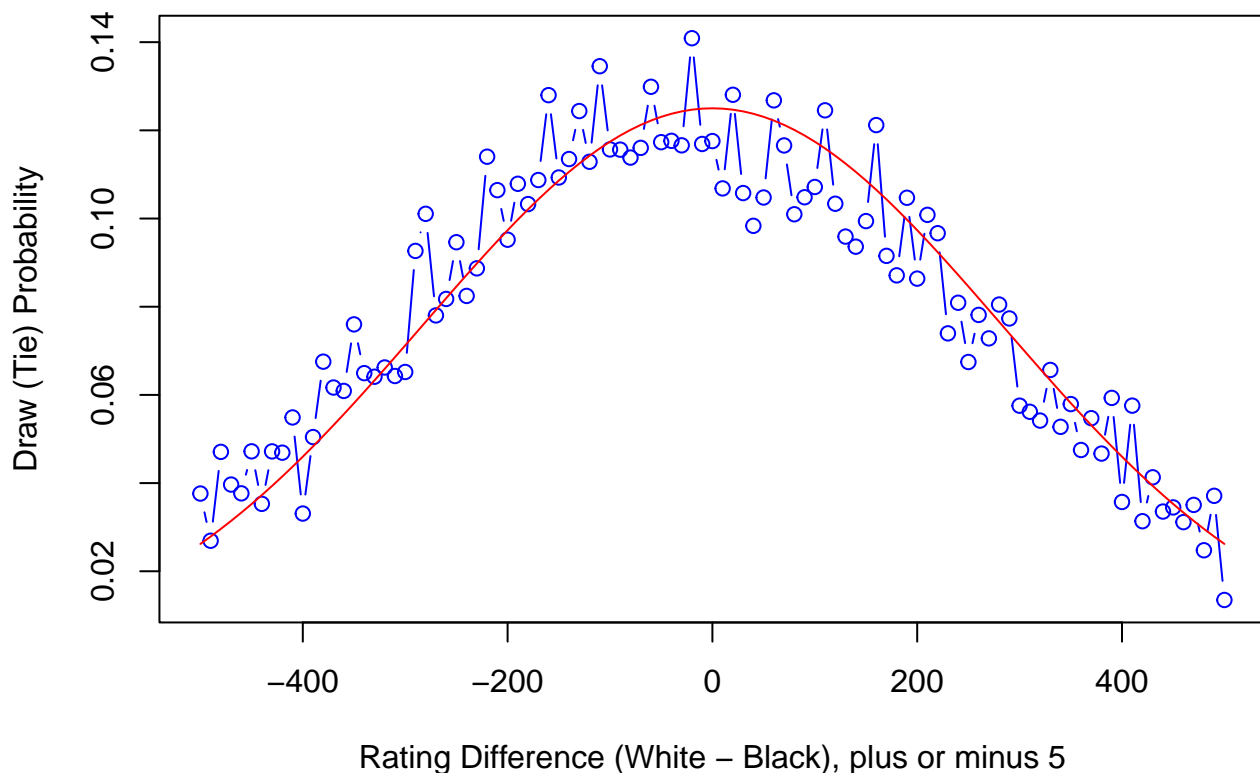
So, in our analysis below, we use the formula $(**)$ for the probability of a tie. It then follows from the above that the probability of a win is given by

$$W \;=\; S - \frac{1}{2}\,T$$

with S as in $(*)$ and T as in $(**)$.

Independence versus Hot Hand:

To model probabilities of streaks, another issue is the extent to which different games are independent. There is a long history of statistical debate about "hot hands" in basketball and other sports. It is quite plausible that there would be some "hot hand" or persistence of performance in chess games as well, especially for games played on the same day in rapid succession, and perhaps even against the same opponent.

To investigate this, we examined the 57,421 games played by Hikaru on Chess.com. For each game, we computed Hikaru's "excess score", defined as his actual score (i.e. 1 or 0 or $\frac{1}{2}$) minus his expected score E from $(*)$ (which depends on his rating difference with his opponent in each game). This gives a time series list of excess score for all 57,421 games played by Hikaru on Chess.com, in time order.

For such a time series, we can consider their "autocorrelations" which measure, for each time lag, the correlation between the excess score on games played at that spacing. For example, at lag=1, this measures the correlation of excess score between successive games. (The autocorrelation at lag=0 is always equal to one, since games have perfect correlation with themselves.) For Hikaru's list, the autocorrelations are as follows:

Autocorrelations of Hikaru Excess Scores
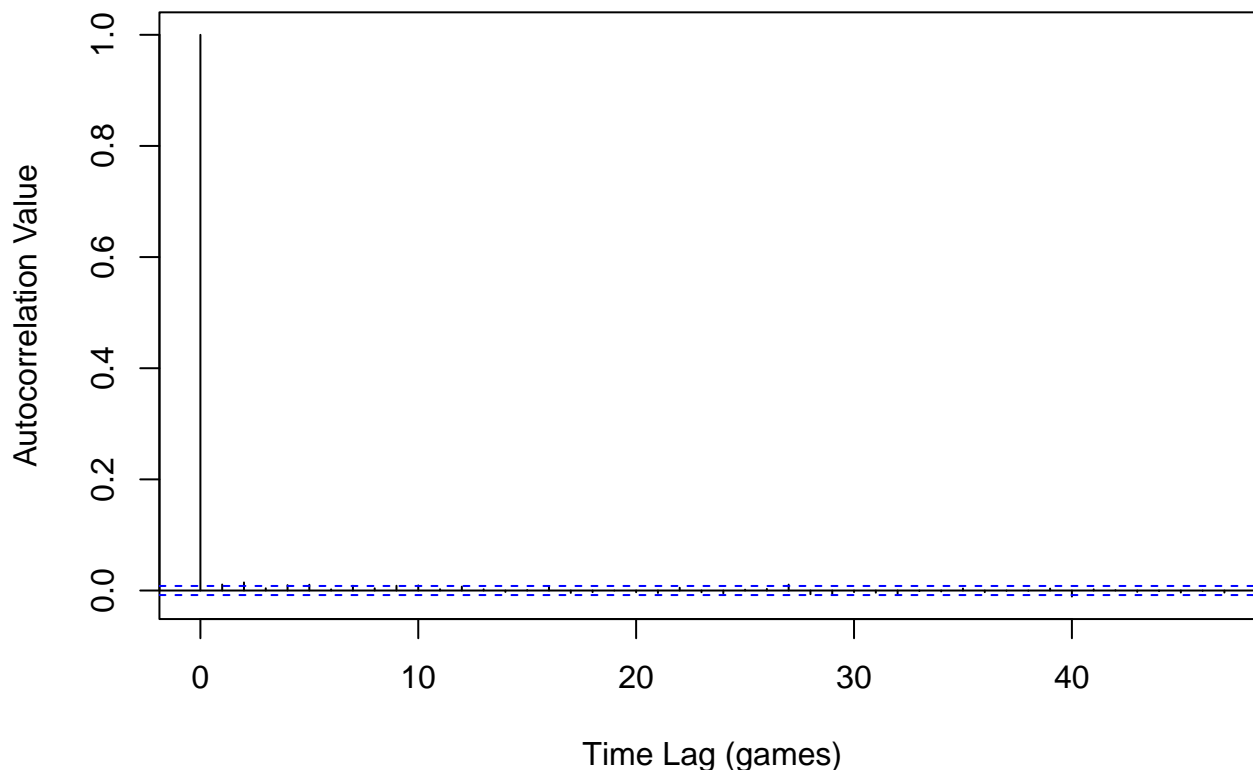
We see from this plot of autocorrelations that (to our surprise) the autocorrelations at all lags (except for lag=0) are all extremely close to zero. This indicates that there is virtually no correlation between Hikaru's excessive scores on successive games. That is, for these games at least, there is no overall evidence of a "hot hand", so they can be regarded as all being independent. We use this assumption in our analysis below.

Hikaru Winning Streaks – Identification:

Next, we investigate winning streaks in the Hikaru game data.

Hikaru is recorded as playing a total of 57,421 games on Chess.com over the date range 2014-01-06 to 2024-07-14, primarily at time control 3m+0s (i.e., three minutes each for the entire game; 35,449 games) or 1m+0s (15,569 games) or 3m+1s (3,310 games). We combine all of these games together, in time order, to determine streaks. (It is also possible to separate out the games played at specific time controls, and/or against highly-rated opponents only. We have investigated this too, and the results are similar or less streaky; see below.)

To define a "streak", we have to decide how to handle draws. At the "pure" extreme, streaks consist solely of wins, while any draw or loss ends it. (We did investigate such

"pure" winning streaks, and again the results are similar or less streaky compared to the below.) At the other extreme, wins or draws both continue a streak, while only a loss ends it. (This is a very loose definition, allowing many draws in a row to constitute a major "streak", which seems inappropriate.) As a compromise, since the most controversial of Hikaru's streaks involved just one draw, we use the "in between" definition that a streak consists of a sequence of games with no losses and at most one draw. That is, a single draw continues a streak, but a second draw (or any loss) ends it.

With this definition, Hikaru has a total of 8,069 streaks (including some overlapping ones). Now, most of these are very short "streaks"; indeed 1,302 of them consist of just a single game. However, quite a few of them are reasonably large. Indeed, 226 of them are at least 30 games, and the largest are of lengths 121, 114, 107, 103, and 101.

Hikaru Winning Streaks – Probabilities:

Just because a streak is long, does not necessarily mean that it is unlikely. We define the "raw" probability of each streak as the probability of his either winning all of those games (if he did win them all), or of either winning all or winning all but one and tying the other (if he did win all but one and tie one). This probability depends on the rating differences, according to the formulae (∗) and (∗∗), using the actual rating difference of the players at the time of each game (to account for changes in player ability over time). We assume independence between games as discussed above; any persistence or "hot hand" assumption would only make the streaks more likely. We can then compute the "raw" probability of that specific streak on those specific games.

Even some very long streaks are not particularly unlikely. For example, Hikaru's streak of length 121 began with his game number 20,940, and took place on 22 December 2018 (except for the final game), with opponents having a mean rating of just 1,579 (compared to his rating of over 3,000 during that same period). His probability of scoring at least 120.5 on those 121 games then works out to 1/8.9, which is not unlikely at all. (Streaks which are very long but not unlikely might be related to Hikaru's apparent practice of "farming", i.e. intentionally seeking out lower-rated players in an effort to increase his own rating [12].)

To avoid streaks which are short, or not particularly unlikely, we need to focus on those which are striking in some sense. To be specific, we define a streak to be "notable" if it consists of a total of at least 30 games, and has raw probability less than one chance in 500. The following table shows all of Hikaru's streaks which are "notable" by this definition:

Hikaru Notable Streaks (length 30+, and prob 1/500+):

line	streaknum	enddate	startgame	length	score	expected	1/prob
1	589	2016.04.06	7027	54	54	48.6	1766
2	2155	2018.07.01	18184	44	43.5	38	875.9
3	2414	2018.11.12	19665	40	39.5	33.9	1083.7
4	2415	2018.11.12	19666	57	56.5	49.2	9452.1
5	3734	2020.01.25	28227	30	29.5	23.9	1345.1
6	3917	2020.03.12	29340	41	40.5	33.3	11570.6
7	4465	2020.05.31	32790	61	61	55	4849.8
8	4551	2020.06.30	33483	53	52.5	47.1	808.1
9	7388	2023.11.17	51857	46	45.5	40	829.6
10	7770	2024.03.12	55162	35	34.5	29.5	568.4

<u>Hikaru Notable Streaks – Theoretical Analysis:</u>

As can be seen from the above table, the most controversial streak, of length 46 ending on date 2023.11.17 (line 9), is not too far out of the ordinary. It has probability about one chance in 830. Indeed, a sequence of 57,421 games has about $57,421/46 \doteq 1,248$ different non-overlapping independent chances to achieve a streak of that length, so finding one with probability 1/830 is actually very likely.

Of the other streaks, just two are considerably less likely, namely lines 4 and 6 with probabilities 1/9,452 and 1/11,571 respectively. So, we consider them next.

The streak on line 4, with probability 1/9,452, consisted of 57 games during the period Nov. 10–12, 2018, at time controls 3m+0s (37 games) or 1m+0s (20 games). (As an aside, the previous streak on line 3 largely overlaps with this one, beginning with a draw one game earlier, and thus ending earlier upon the next draw.) This streak was played against a total of eight different opponents (none more than ten games each). The mean opponent rating was 2,786, or about 315 less than Hikaru's rating which hovered around 3,100.

Now, out of 57,421 games total, there are still over 1,000 different independent opportunities to establish a streak of length 57 (and much more if overlapping, non-independent intervals are considered). So, as a first approximation lower bound, suppose there are 1,000 independent opportunities to establish a streak, each of independent probability 1/9,452. Then the probability of achieving such a streak would be given by

$$1 - \left[1 - \frac{1}{9,452}\right]^{1,000} \doteq 0.100398 = 1/10.$$

That is, under this independence approximation, the probability of achieving such a streak over the course of 57,421 games is about one chance in ten. That is not particularly surprising, and does not reach the usual 0.05 level for statistical significance. (And remember that this is just a *lower bound*, ignoring overlapping opportunities; see the next section.)

The streak on line 6, with probability 1/11,571, consisted of 41 games during the period March 10–12, 2020, all at time control 3m+0s. This streak was played against five different opponents, including Bigfish1995 (13 games) and ToivoK3 and wonderfultime (10 games each). The mean opponent rating was 3,008, which on average was nearly 250 less than Hikaru's rating which had climbed to around 3,250 by this point.

Out of 57,421 games total, there are about 1,400 different independent opportunities to establish a streak of length 41 (again ignoring overlapping opportunities). So, we consider the approximation that there are 1,400 independent opportunities to establish a streak, each of independent probability 1/11,571. Then the probability of achieving such a streak would be given by

$$1 - [1 - \frac{1}{11,571}]^{1,400} \doteq 0.11396 \doteq 1/8.8\,,$$

That is, under this lower bound approximation, the probability of achieving such a streak over the course of 57,421 games is about one chance in 8.8. That is not very surprising, and again does not reach the usual 0.05 level for statistical significance.

Hikaru Monte Carlo Simulation:

The above discussion indicates that Hikaru's individual win streaks are not particularly surprising. However, the approximate probabilities computed above are just lower bounds, since they do not take into account the additional possibilities of long streaks in overlapping (and hence not independent) game sequences.

To analyse this further, we conducted a Monte Carlo (random) simulation. Specifically, using the actual player ratings for each of Hikaru's 57,421 games, we simulated fresh independent results using the probabilities of wins and ties from (∗) and (∗∗). We repeated this simulation 100 different times, each time recording the smallest individual streak probability (i.e. the largest 1/probability), and also the total number of "notable" streaks as above (i.e. length at least 30, and 1/probability at least 500).

The distribution of the largest 1/probability in those 100 simulations, together with Hikaru's observed two largest 1/probability values (11,570.6 and 9,452.1, respectively) in red, is as follows:

Hikaru Streak Probability Monte Carlo Samples
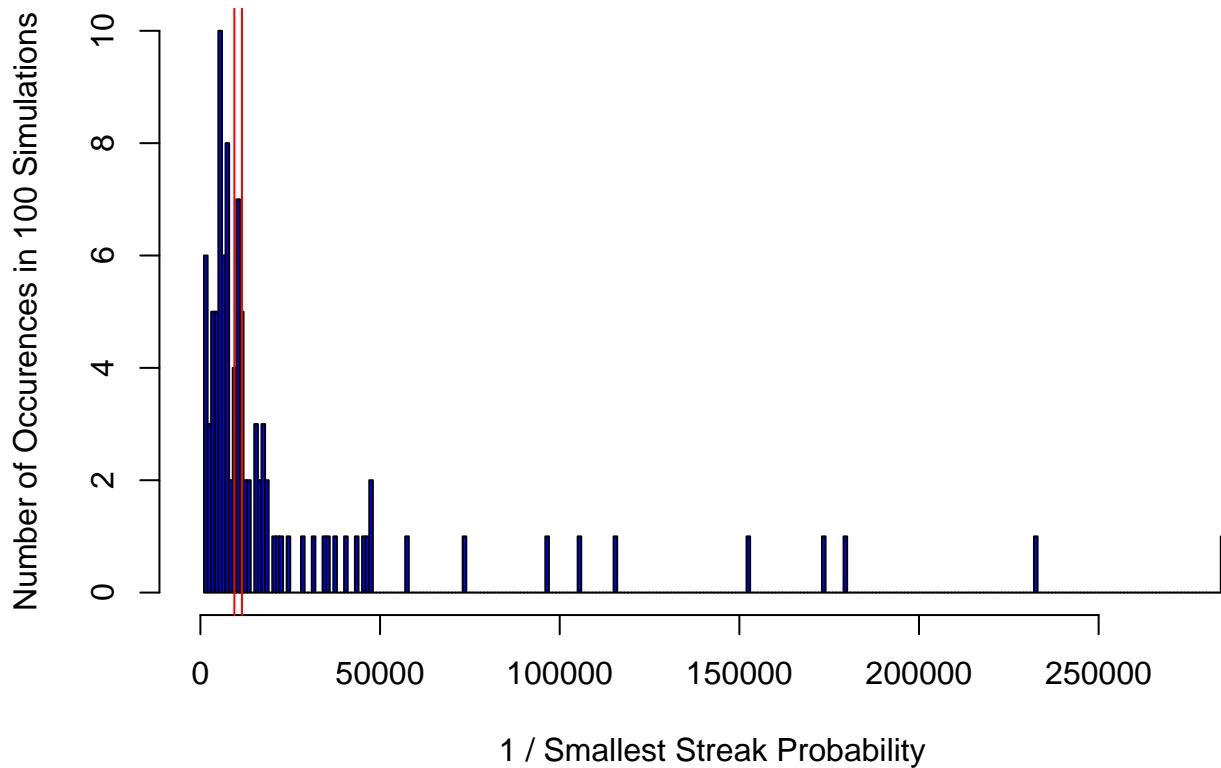
We see from the graph that, while the observed values 11,570.6 and 9,452.1 are larger than many of the simulated maximum 1/probability values, there are also many simulated 1/probability values which are much larger than that. Indeed, the largest simulated 1/probability value is over 284,000, and the mean simulated 1/probability value is over 26,000, and even the median simulated 1/probability value is 10,462 which is close to Hikaru's 11,570.6 value. In fact, in 43 of the 100 simulations (nearly half), the least likely streak is less likely than the observed 1/11570.6 one. This further confirms that Hikaru's least likely streaks are not surprising over such a long collection of games.

Similarly, the distribution of the *number* of "notable" streaks (i.e. length at least 30, and 1/probability at least 500) in each of the 100 simulations, together with Hikaru's observed value of 10 in red, is as follows:

## Hikaru # Notable Monte Carlo Samples
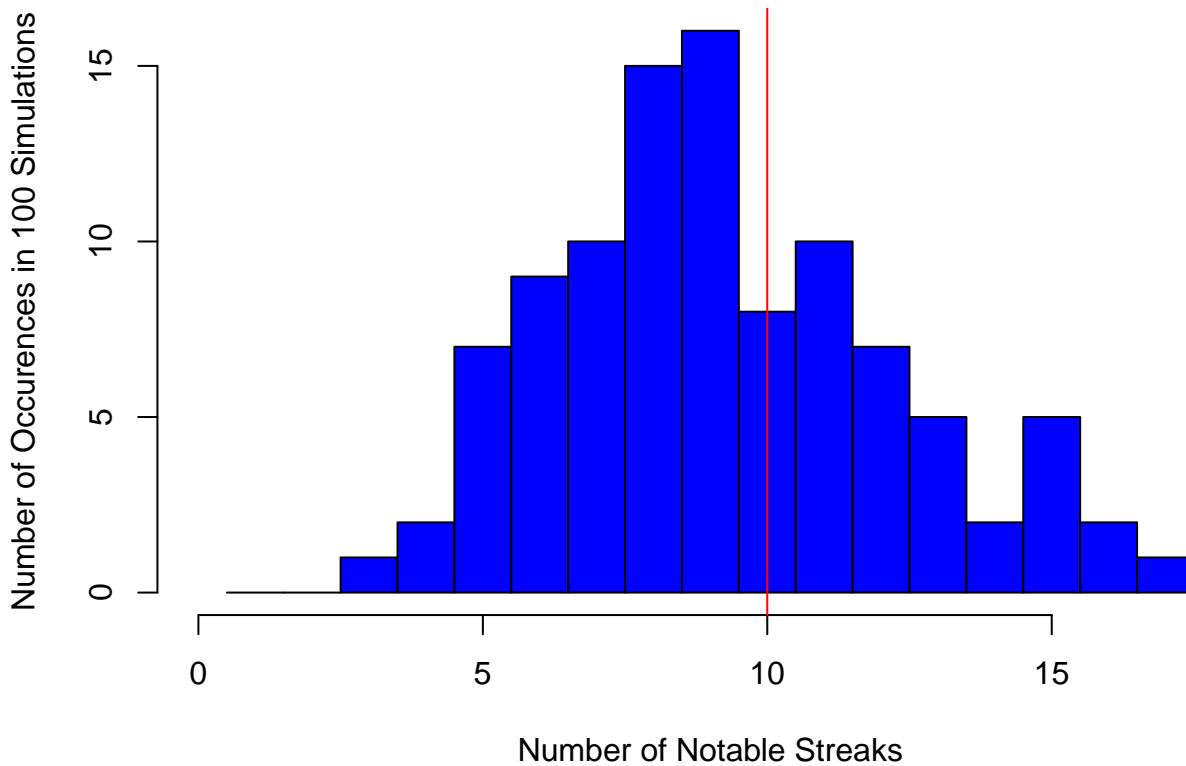
This graph shows that Hikaru's actual number 10 of notable streaks is quite typical for the simulations, which have a mean of 9.25 and median of 9. Indeed, in 40 of the 100 simulations, the number of notable streaks was equal to or greater than the observed 10. This confirms that Hikaru's number of notable streaks is also not surprising.

Second-Least-Likely Streaks:

As another test, we note that Hikaru had *two* streaks whose raw probabilities were quite low, namely 1/11570.6 and 1/9452.1. As discussed above, each of these streaks on its own turns out to be not at all surprising. However, this raises the question of how unlikely it would be to have two such streaks which are *each* so unlikely.

To test this, we ran another Monte Carlo simulation, again simulating the possible outcomes of Hikaru's 57,421 games, but this time looking at the raw probability of the *second* most unlikely streak. Once again, we repeated this simulation 100 times. The resulting distribution, together with Hikaru's observed value of 1/9452.1, appears as follows:

Second–Most Unlikely Streak Monte Carlo Samples
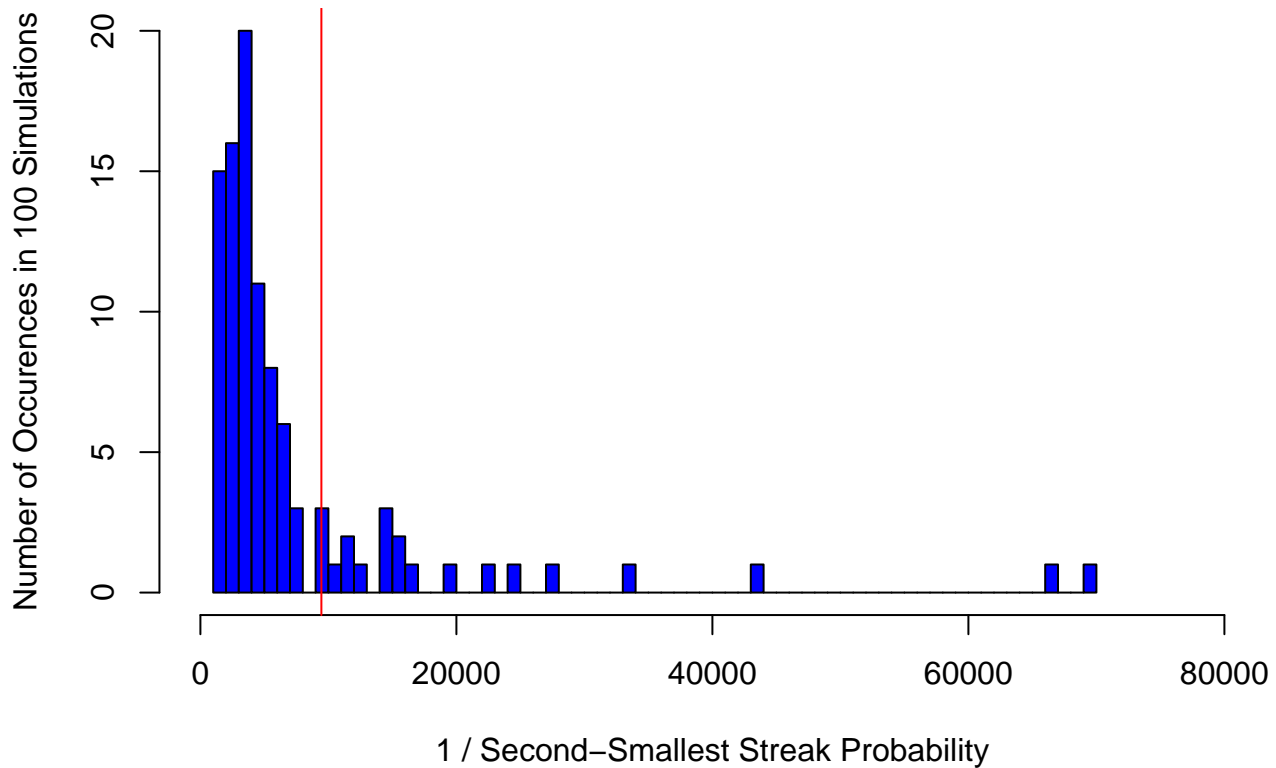
This graph indicates that Hikaru's second-most-unlikely streak probability of $1/9452.1$ is slightly less likely than typical. Indeed, only 18 of the 100 simulations had second-most-unlikely streaks which were less likely. Nevertheless, even 18% is still quite a large fraction, much larger than the 0.05 required for statistical significance. Indeed, some of the simulated second-most-unlikely streaks were considerably less likely, with two of their probabilities less than $1/65,000$. This analysis indicates that Hikaru's second-most-unlikely streak was slightly less likely than expected, but still well within the usual range of statistical fluctuation.

It is possible to continue such analyses to additional streaks. However, Hikaru's third-least likely streaks (and beyond) are less surprising (with probabilities $1/4849.8$ and larger), so they provide less evidence. Indeed, a fresh Monte Carlo simulation looking for third-least-likely streaks found that 45% of the time they are less likely than $1/4849.8$, so this is not rare at all. And his fourth etc streaks would be even less rare by this measure. That is, Hikaru's two least likely streaks are the only two with very small raw probabilities which are potentially surprising and thus in need of investigation.

Restricting to just a Single Time Control:

The response [10] vigorously objected to my combining games at different time controls when investigating streaks. So, I now consider what would change if we restrict solely to games with 3m+0s time control (Hikaru's most common time: 35,449 games, including the entire controversial 46-game streak). In that case, the list of "notable" streaks (as defined above) becomes the following:

Hikaru 3m+0s Notable Streaks (length 30+, and prob 1/500+):

line	streaknum	enddate	startgame	length	score	expected	1/prob
1	8	2014.05.14	45	46	46	40.5	2143.3
2	1189	2018.07.01	8757	44	43.5	38	875.9
3	2400	2020.01.25	16508	52	51.5	46.5	680
4	2557	2020.03.12	17551	41	40.5	33.3	11570.6
5	3063	2020.05.31	20723	61	61	55	4849.8
6	4575	2023.07.10	31607	47	46.5	41.1	746
7	4675	2023.11.17	32499	46	45.5	40	829.6

Comparing this with the previous table, we see that several of the previous streaks remain, including the least likely one (now line 4), and the most controversial one (now line 7). (However, the second-least-likely one disappears.) And, in a fresh Monte Carlo simulation of 100 random results for just these 35,449 games, 29% of them have a streak which is less likely than the 1/11,570.6 lowest probability achieved by Hikaru. So, Hikaru's notable streaks, restricted to just 3m+0s games, are still not at all unlikely over his full history.

Related to this, the above fits of the expected score and draw probabilities to average game data were also done combining the seven data files at all time controls (a total of 293,047 games). If we restrict to just the 3m+0s games (a total of 131,445, or about 45% of them), then the fit of the score formula (∗) to average scores becomes:
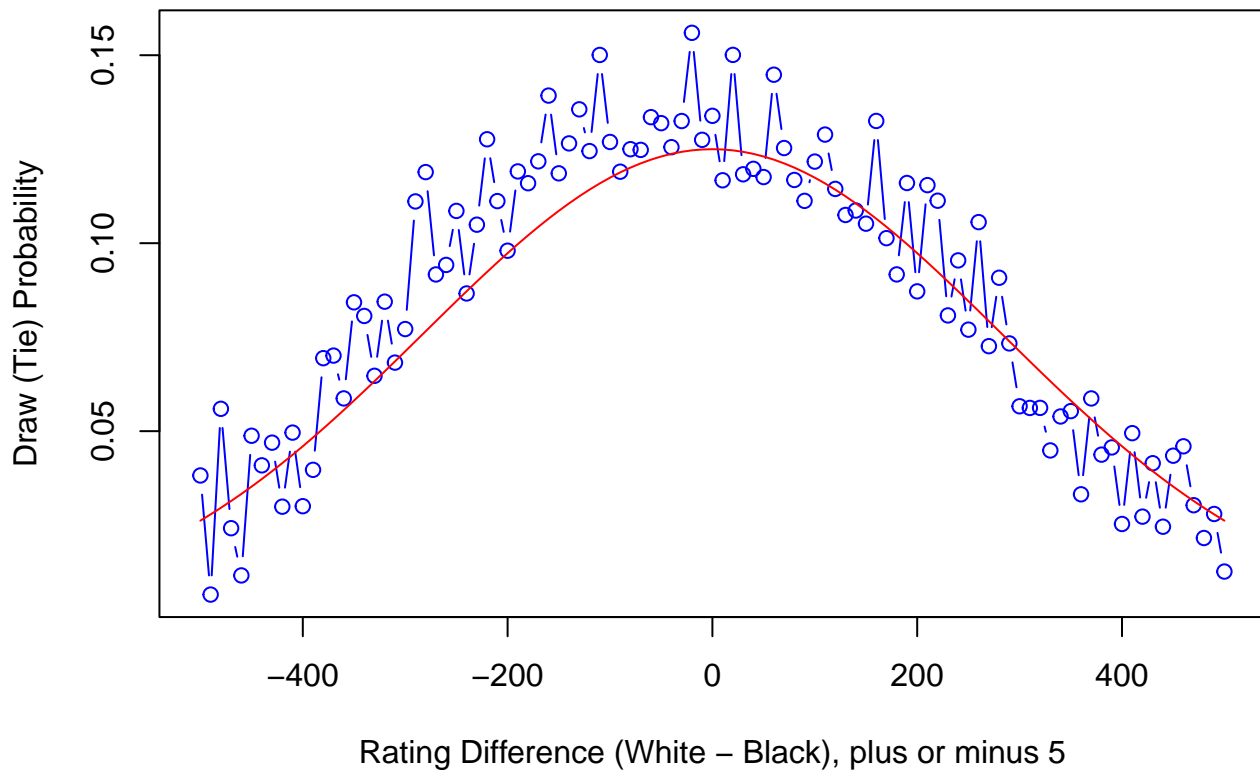
Average 3m+0s Scores versus Adjusted Expected Scores
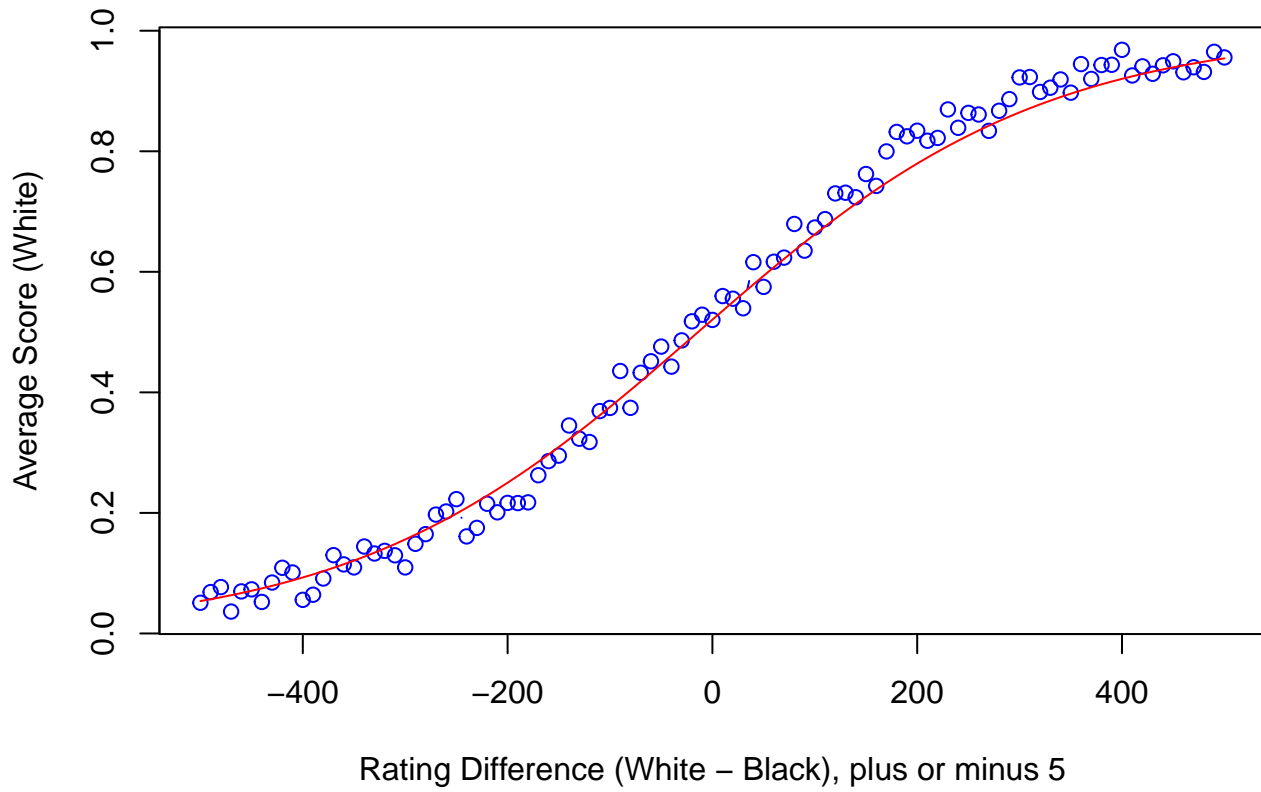
This is nearly as good a fit of (∗) as before, with just *slightly* more excess at differences around +200, but still quite a reasonable approximation.

Meanwhile, the fit for draw probabilities (∗∗) for just the 3m+0s games becomes:

Probability of a 3m+0s Draw (Tie)

This is a slightly worse fit of (∗∗) than before, since the 3m+0s games have a slightly higher draw rate than the data overall. So it could be adjusted slightly, but this would make only small changes to the probabilities, not nearly enough to effect the overall conclusions.

In summary, restricting to just the single most common 3m+0s time control makes only small changes to the fit of the (∗) and (∗∗) formulae, and does not show any additional evidence of unlikely winning streaks.

Conclusion:

This statistical analysis indicates that Hikaru's online chess winning streaks are not particularly surprising. His recent controversial streak of length 46 is well within expected levels. His two most surprising streaks are of length 57 in 2018, and of length 41 in 2020. Although their raw probabilities are each about one chance in 10,000, the probability of observing each such streak over the course of so many games is still shown to be above 10% based on independent non-overlapping opportunities alone, and about 43% in Monte Carlo simulations, and hence not unlikely. Having *two* such notable streaks is somewhat less likely, but still occurs about 18% of the time, well within usual statistical variability.

Of course, every statistical analysis requires making some choices regarding definitions, scope, etc. I believe that the choices in this report are all fair and defensible, consistent with the available data, and lead to accurate conclusions. Furthermore, various alternative choices (such as restricting to a single time control) lead to similar results, with no indications that other reasonable choices would lead to significantly different conclusions. Hence, the final conclusion, that the streaks observed in Hikaru's Chess.com record are fairly typical given the player ratings over Hikaru's long record of games, appears to be quite robust.

In closing, contrary to the allegations in [10], I seek only the truth, and am not at all averse to statistically identifying suspicious behaviour it arises (see e.g. my work in the high-profile lottery retailer scandal [17]). So, I would be very happy to uncover suspicious chess behaviour if I could. However, I do not see any evidence of suspicious behaviour in Hikaru's online chess streaks.

References

[1] G. Keener, Chess is Booming. Published in the New York Times, 17 June 2022. Available at: https://www.nytimes.com/2022/06/17/crosswords/chess/chess-is-booming.html

[2] What do I need to know about Fair Play on Chess.com? Available at: https://support.chess.com/en/articles/8568369-what-do-i-need-to-know-about-fair-play-on-chess-com

[3] Chess.com Fair Play Report: Estimating Cheating Rates in Titled Tuesday. 18 July 2024. Available at: https://www.chess.com/blog/FairPlay/chess-com-fair-play-report-estimating-cheating-rates-in-titled-tuesday

[4] D.L. McClain, Elite Chess Players Keep Accusing Each Other of Cheating. Published in the New York Times, 25 December 2023. Available at: https://www.nytimes.com/2023/12/25/crosswords/chess-hikaru-vladmir-kramnik-cheating.html

[5] Vladimir Kramnik interview with GothamChess. July 2024. 35 min. Available at: https://www.youtube.com/watch?v=yDeQ-CmeZaU

[6] Wired Q&A with J.S. Rosenthal. February 2022. 17 min. 2.2 million views. Available at: https://www.youtube.com/watch?v=QW3KRaz4aI4

[7] Direct email from Erik Allebest (Chess.com CEO), 15 July 2024.

[8] J.S. Rosenthal, Probabilities of Streaks in Online Chess (original version). 9 August 2024. Archived at: http://probability.ca/jeff/ftpdir/chessstreaksorig.pdf

[9] T.J. Svensen, Professor Finds Nakamura's Winning Streaks Statistically Normal, Kramnik Dismisses It. Chess.com article, 30 August 2024. Available at: https://www.chess.com/news/view/nakamura-winning-streaks-statistically-normal-professor-says

[10] V. Kramnik, Another research full of mistakes. 30 August 2024. 39 min. Available at: https://www.youtube.com/watch?v=qEsUXqnuDa8

[11] J.S. Rosenthal, Probabilities of Streaks in Online Chess (addendum version). 5 September 2024. Archived at: http://probability.ca/jeff/ftpdir/chessstreaksadd.pdf

[12] Hikaru Nakamura, Prof. of Stats Reports – Kramnik: "Lies!!". 3 September 2024. 21 min. Available at: https://www.youtube.com/watch?v=bfn9gAEY0Ck

[13] K. Bobyrev, Analyzing long win streaks in online chess. Blog post, 29 November 2023. Available at: https://kirillbobyrev.com/blog/analyzing-long-win-streaks/

[14] Chess.com, Regarding Recent Accusations. Posted 29 November 2023. Available at: https://www.chess.com/blog/CHESScom/regarding-recent-accusations

[15] S. Maharaj, N. Polson, and V. Sokolov, Kramnik vs Nakamura or Bayes vs p-value. Preprint, 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4648621

[16] M.E. Glickman, The Glicko rating systems. Described at:
http://www.glicko.net/glicko/glicko.pdf
and http://www.glicko.net/glicko/glicko2.pdf

[17] J.S. Rosenthal, Statistics and the Ontario Lottery Retailer Scandal. Chance Magazine **27(1)**, 2014. Available from: http://probability.ca/lotteryscandal/