

Motivation

The anti cheating committee has been receiving a fair few allegations of cheating over the first two weeks of the 2020 event.

A lot of these complaints seem to be based on a statistical estimate of playing strength that chess.com provides called the Computer Aided Precision Score (CAPS).

What is CAPS

CAPS is an algorithm that attempts to evaluate a player's chess strength based on moves rather than results.

Chess.com's David Pruess wrote an article describing their CAPS tool that can be found at <https://www.chess.com/article/view/better-than-ratings-chess-com-s-new-caps-system>

First of all I would like to state that this is definitely a **fun way of getting some intuition of how well (badly) you are playing**. I run pretty much all my chess.com games (ok, mostly just my wins...) through it.

However CAPS is DEFINITELY **not very good evidence that your opponent is a cheat** (seriously - if you only take one thing from this article then it is these two (:D) comments in bold)

What is Wrong with CAPS

First of all I wish to draw your attention to this statement from the guy whose tool this is (see article I linked to) and pay particular attention to the words that I have underlined:

“Below is a graph that shows what someone's CAPS might be at the given rating levels currently used by the chess world to measure someone's chess skill.

Note that it might not mean you, as a 1400-rated player, will always have a CAPS of 79.189 in your games... However, with enough data (games reviewed) your overall CAPS is likely within the target range.”

There are two important points covered in the above quote.

Possibly the most important one is encapsulated in the “**with enough data (games reviewed)**” part. It is important to remember that CAPS is essentially a statistic and, among other things, for a statistic to be reliable you need a decent **sample size**.

Now it’s a bit debatable what a decent sample size is in this case (200 moves? 200 games?) but it is certainly not going to be one game (or two games) except in perhaps the most egregious cases.

The second point is a bit more subtle. Look at all those phrases I have underlined... “might be”, “likely to be”... “within the target range” [without specifying what this target range is at all - horrifying].

In my day job as a data scientist I would not get away with using this kind of language when talking about the performance of any algorithm I deployed. The reason for that is that these are all weasel words (https://en.wikipedia.org/wiki/Weasel_word) and there is no quantification of uncertainty or even performance whatsoever. Reading that chess.com article would you be able to state that the algorithm they use is right 99% of the time? Obviously not. What about more than 75% of the time? 60%? If you read their article then the only claim they have made that puts any sort of lower bound on the performance of their algorithm is the word “likely”. Ok great. They are at least (arguably) claiming >50% accuracy. Not a good basis for accusing someone you don’t know anything about.

I strongly suspect the reason they have put those words in is because this is just a very rough estimate and they think (hope!) that it is going to give a ballpark figure, averaged over a decent number of games, in a majority of cases. Nothing wrong with that - provides lots of fun - not something to accuse people off the back of.

Specific Issues With Applying CAPS Scores To Small Numbers of Games

The following is by no means an exhaustive list:

- CAPS hugely overestimates people's scores when the position is completely dead-drawn for a lot of moves (shuffles in opposite coloured bishop endgames)
- CAPS attempts to adjust for position complexity but it fails - it still overestimates people's playing strength in dull games.
- CAPS often hugely overestimates the winner and similarly underestimates the loser if the game is short
- CAPS doesn't know how much of the game is theory/preparation
- CAPS will give a much better score for wins than for losses (obviously, but bear that in mind when you are looking at a game you have lost and that therefore the person you are accusing has won...) - the score is meant to work by looking at wins, draws and losses together over large numbers.
- CAPS often gives better scores if you are playing against players much lower rated than yourself - it's easier to look good if you are playing someone much worse than you.
- People do play better than their rating sometimes - especially in one off games.
Otherwise there would be no point in playing games with big rating differences.

In conclusion, CAPS probably is quite good at ruling out cheating in cases where it gives a very low score but I wouldn't use it as a basis for accusing someone of being a cheat if it has given a high score. Accusing someone of cheating based only on statistical evidence is a very complex endeavour