The official website of the United States Chess Federation | Press | Advertising | About the USCF | Contact Us

Search | Login or join now to view members-only content | Member Login | Join Today

Join or Renew
New to Chess?
USCF Shop
Chess Life Magazine
Chess Life Online
Top Player Bios
Tournaments/Clubs
Ratings
Member Services Area
Scholastics
U.S. Chess Trust 501(c)3

New site! Report problems at support@uschess.org
uschess.org > Chess Life Magazine > A Conversation with Mark Glickman

A Conversation with Mark Glickman

Josh Kuchinsky

One of the most authoritative lay articles ever published about chess ratings is a 43-page survey, modestly titled, "Chess Rating Systems," written for the third, and final, issue of the American Chess Journal. The year was 1995. The writer was a young, 30-year-old Mark Glickman, only two years removed from his graduate school days at Harvard, but already a ten-year veteran and chairman of the USCF Ratings Committee—a position he retains to this day.

Glickman is considered one of the world's leading authorities on chess ratings. He has been instrumental in designing, and is responsible for maintaining, the integrity of the current USCF rating system, and has published, in peer-reviewed journals, sophisticated, modern improvements to Elo's original theory—improvements so significant that his new systems, Glicko and Glicko-2, have become standards used by numerous online gaming communities and, just recently, the Australian Chess Federation. Glickman works as an Associate Professor at Boston University and as a Senior Statistician at one of the Veterans Administration research centers, The Center for Health Quality, Outcomes and Economics Research, where he performs statistical modeling on healthcare related issues. If you visit his website, his curriculum vitae scrolls endlessly, long past one page.

This is the standard rating formula: **R new = R old + K(Wactual – Wexpected).**

It is simply an expression of arithmetic stating that a new rating equals an old rating, changed by the difference between actual results and expected results, multiplied by a scaling factor, K. Performing this arithmetic is easy. Understanding the qualitative ideas behind this arithmetic, the ideas striving to model the uncertainty of playing like a human, is not so easy, however. In other words, what are the appropriate values for K and Wexpected? Finding an answer to this question is what propels ratings-related research. It is what keeps Mark Glickman busy.

There has been very little mainstream ratings-related chess literature published within the last eleven years, especially literature focused on the USCF system. This has, perhaps, spawned qualitative misconceptions about what a chess rating is supposed to represent. For instance, we often hear chess players say something like, "I just beat a 1249." This statement carries a lot of presumptions, the most common one being that a rating represents some absolute value of chess strength rather than an estimation of human performance. In this interview, Glickman, in his characteristically enjoyable and precise style, explains how human variable

Top Stories
Kamsky Wins but the Cat Makes News
World Open web exclusive!
Heated Play at Miami Open
Vukcevich, the Solver!
The Chess-Go Continuum
Counterplay
The Neutron Bomb Middlegame
Across the Board

Archives
September 2006
August 2006
July 2006
June 2006

performance is modeled. Along the way, he covers many details about the current state of the USCF rating system. If there is one message to take away, it is this: Chess players don't play like numbers—they play like people.

Howard Goldowsky: What led you to become a member of the Ratings Committee and eventually its chair?
Mark Glickman: Back in 1985, when I was in my senior year at college, I had begun studying mathematical models for measuring chess strength. I had become particularly interested in how to calculate ratings for unrated players, using some of the statistical methods I had been learning at that time. In fact, I had written my senior-year thesis on the development of a provisional rating system. I thought some of the ideas I was working on might end up being useful to the USCF, so at the U.S. Amateur Team East, in 1985, I spoke with Steve Doyle (the USCF president at the time) about becoming a member of the ratings committee. I was appointed to the committee shortly afterwards. My method for computing provisional ratings was never actually incorporated into the rating system, though the basic principles underlying my approach have been adopted and are being used in the current provisional rating algorithm. After being an active member for several years, I was finally appointed chair of the Ratings Committee in 1992. I have remained chair ever since (sharing the position with Frank Camaratta for a number of years).

HG: What type of work does the Ratings Committee do?
MG: The USCF Ratings Committee is essentially a group that advises the USCF office and the USCF Executive Board (EB) on ratings-related issues. The committee currently consists of chess players and organizers with a strong quantitative background, as well as professors in statistics, computer science, and economics. It is rare that we reach unanimous agreement on any particular issue, but the discussions can be lively and thought-provoking. The most common occurrence is that the EB charges the Ratings Committee with particular tasks, and then we respond by making recommendations after a sufficient period of review. These tasks tend to involve issues concerning special cases of the rating formulas. For example, the committee has, in the recent past, been asked to evaluate whether a separate rating system should be established for blitz chess (with time controls ranging from G/9 through G/3), or whether blitz games should be handled under the current Quick Chess rating system. Another, more interesting, example is how the USCF rating system can be adapted to update the ratings of players who compete in foreign FIDE tournaments, that is, tournaments that are not normally rated by the USCF. Besides taking an active part in general ratings discussions with the committee, my role is to present and lead the discussion on the ratings issues charged to us by the EB, and then to serve as the liaison between the EB and the USCF office.

HG: Your interest in ratings actually extends beyond the USCF. In fact, you've made ratings an important part of your career. Could you please elaborate on this and also describe some of your ratings-related research?
MG: While much of my normal academic work has to do with statistical aspects of designing and analyzing healthrelated studies, a major part of my work involves research into ratings issues. My original focus on ratings research was the development of the rating systems, Glicko and Glicko-2. More recently, however, I have become very interested in methods and principles for determining player pairings within tournaments. The amount of serious statistical research in this area has been surprisingly small, and I think the field is wide open to make interesting contributions. I have also been involved in a number of consultations with gaming organizations (typically online gaming) to help them develop rating systems for online play. These are usually small projects, often helping organizations implement versions of Glicko and Glicko-2.

HG: Could you briefly describe the Glicko system, and explain, qualitatively, how it is different than the current USCF system?

MG: The Glicko system came about as a consequence of my PhD dissertation. My dissertation developed statistical models of playing strength, recognizing that players' strengths could be time-varying. It approached the problem from first principles and applied the methods not just to chess, but also to National Football League football game outcomes. The methods in my dissertation required an enormous amount of computation to be of practical use, and so I developed the Glicko system to address this computational burden.

To appreciate the basic concepts of the Glicko system, it's worthwhile to understand how a statistician might think about ratings. In general, a statistician makes a distinction between an estimated rating and a true, underlying rating called the rating "parameter." The rating that appears on your Chess Life mailing label or in any published form is your estimated rating. The rating parameter, on the other hand, which is never actually known, is the real description of one's strength. For example, when someone claims he is underrated, what he really means is that his estimated rating is less than his rating parameter; that is, his published rating is lower than his true strength. The goal of a good rating system is to produce rating estimates that are close to the corresponding and unknown rating parameters.

To be even more specific, a rating parameter really only represents one's average playing strength. Suppose that a player's rating parameter was 1800 (even though this would never be known), when this player sits down to a game, his strength for that game might be 1780, it might be 1830, or it might be some other value not too far from his average strength of 1800. The point is that a player's strength varies a bit from game to game around his average. This average is the rating parameter. A statistician's simplistic view of a game of chess, then, is that the player who produces a higher strength for that particular game will be the winner. A player who has a higher rating parameter will tend to produce higher game strengths than an opponent with a lower rating parameter. So, if two players compete and the rating parameter for the first is 1750, and for the second it is 1825, then it is more likely that the 1825 player will produce a playing strength for that game which is higher than that of the 1750-player, though there is still a chance that the 1750- player's displayed strength will be higher.

The main qualitative difference between the Elo and Glicko systems is that the Elo system really doesn't make a distinction between rating estimates and rating parameters, and the Glicko system does. Specifically, the Glicko system recognizes that the estimated rating may not be close to a player's rating parameter, and then quantifies this uncertainty in what has been termed by others as the "rating deviation" (RD). For example, a player who has only played four tournament games is bound to have an estimated rating that is not a precise reflection of his true strength (large RD), whereas a player who competes every weekend will have an estimated rating that is fairly accurate (small RD). It's interesting to know that the Elo system can be viewed as a special (but untenable) case of the Glicko system, namely when everyone's RD is exactly zero. That corresponds to the situation where everyone's estimated rating is equal to their rating parameter. To me (and probably to most statisticians), this is a nonsensical assumption to make for a rating system.

HG: How, then, is Glicko-2 different than Glicko?

MG: In the late 1990s, I became interested in the statistics of financial investments. At that time, I noticed some statistical models that were becoming popular for tracking the price of financial products, called "stochastic volatility" models. The basic idea is that the price of an investment might fluctuate over time in a fairly controlled manner, but every once in awhile the price might jump (often downward!) by a magnitude that would appear surprising or unpredictable. The question would be whether this jump was just a fluke, in which case the price would move back to the original price over a short period of time, or whether this jump signified something more permanent.

It occurred to me that this phenomenon of a sudden jump could happen in chess strength. For example, a player could have an off-day, so it seems worthwhile for a rating system to recognize that a bad performance might be only temporary. Or a young player could be improving very quickly and have results that indicate that he is much better than his current rating estimate. These types of situations motivated me to borrow the notion of If there is one message to take away, it is this: "stochastic volatility" and incorporate it into the Glicko rating system. So, in addition to an estimated rating and RD, the Glicko-2 system calculates a value called the "volatility" for each player. A large volatility means that the player's most recent results are unexpected relative to previous results, and the implication in the calculations is that the RD will increase to reflect extra uncertainty due to not knowing what the player's true strength actually is. In contrast, under the Glicko system, a young, rapidly improving player with exceptional results, who plays a lot of games, might actually see his RD go down. I was delighted to learn recently that the Australian Chess Federation adopted Glicko-2 for rating over-the-board tournaments. As I understand, the system has been quite successful, and quickly improving players in Australia are tracked well using Glicko-2. I have begun discussions with the USCF Ratings Committee about investigating whether Glicko-2 would be appropriate for the USCF. This is now in the early stages of discussion.

HG: Four years ago, the USCF rating system underwent a major overhaul. Why did this happen?
MG: Back in 1995, if memory serves correctly, we had 100-point rating floors —if your highest established rating was 1734, your rating floor would be 1600 (100 points below your highest rating, rounded down to the nearest multiple of 100). Because so many people were on their rating floor, the EB at that time agreed to create a 200-point floor. This is where it stands now. Coupled with the drop in the rating floor was an increased influx of scholastic players into the USCF, who were improving more quickly than the rating system could track them. Both of these factors, in combination, resulted in what seemed to be rating deflation, where players who were otherwise at stable strength were consistently losing to young underrated players. The EB claimed that the USCF was rapidly losing members who were frustrated by unfair rating decreases, and wanted the Ratings Committee to address this problem. In response, the Ratings Committee developed a substantially revised rating system. The details of the system had been worked out by 1997, but USCF office difficulties prevented its implementation until early 2001.

HG: What are the main differences between the current, more complex USCF rating system and the one it replaced?
MG: Although at their core the old system and the one it replaced are both based on Elo, there are four main differences between them. First, there is a new method for determining provisional ratings. In most cases, the new formulas produce a provisional rating that is identical to the old system; but when an opponent's rating is far from a player's pre-event provisional rating, or when the opponents' ratings are widely dispersed, the new formulas appropriately use information from results against these opponents to produce an updated rating. Also, the new system often uses agebased imputed ratings to rate games between unrated players. Second, the new system incorporates a bonus point mechanism for tracking quickly improving players. When a player's results exceed the expected result beyond a certain threshold, bonus points are added to the ordinary rating gain. The threshold depends on the number of rounds in an event.

Third, the new system features a "sliding-K" to account for more variable player abilities when players have low ratings or when players have not played many tournament games. The value of "K" in the standard rating formula can be thought of as a measure of uncertainty in a player's pre-event rating, intuitively similar to RD in the Glicko system—the higher the value of K, the greater the impact of a tournament result, and the less reliance on the pre-event rating. The revised rating system, therefore, uses values of K that are large when a player is low-rated or when the player has not played in many USCF-rated games to reflect the greater uncertainty in such players' abilities. Finally, the revised rating system uses an iterative procedure to rate an event. Rather than performing a single rating calculation for each player when rating an event, the revised system performs two iterative calculations for previously rated players and three for unrated players. The main benefit to this procedure is that the results of opponents' games are now incorporated into a player's rating calculation.

HG: Some people might argue that the current rating system is too complicated. How do you respond to this?
MG: I realize that the complexity of the new rating system is troubling to many players. But I think this issue should be put into perspective. When asked about the complexity of the rating system, an analogy I like to use is that many conveniences in our lives are complicated, but this fact tends to make them more usable. For example, cars nowadays are much more complex than, say, 20 years ago, but the features in cars that use modern technology (e.g., GPS, airbags, etc.) make them generally better to drive.

A common reason players are often upset about the complexity of the current system is that they cannot calculate their rating updates after completing a tournament. My response to this is that, technically, it wasn't really possible to do so with the simpler system, either. The reason is that the information on a wallchart is typically not the information that is used when calculating official rating updates. The USCF uses the most current ratings at the time of processing tournament results. So, even if the formulas were simple, the likelihood is that your opponents' ratings on the wallchart at a tournament are not the ones used when the USCF calculates rating updates.

I should also hasten to point out that the Ratings Committee has provided a set of approximating formulas that are, essentially, no more complicated than those in the old rating system. Players can use these formulas to estimate their rating updates after tournaments, recognizing that the actual USCF rating update could be a bit different.

HG: Given the complexity of the USCF rating system, how do you diagnose that it's working correctly?
MG: It's very difficult to diagnose whether the rating system is working. Technically, one should collect information on a large random sample of tournament game outcomes, and then analyze whether the results of the games are consistent with what the players' ratings should predict. Unfortunately, this strategy is difficult to implement—the winning expectancy formula is only meant to apply to rating parameters, and not rating estimates. When applied to published ratings, the winning expectancy formula tends to be overly optimistic for the higher-rated player. I published a paper with Albyn Jones, a statistics professor who is also on the Ratings Committee, where this phenomenon was demonstrated.

Given the difficulty in testing the rating system, what the ratings committee has been doing the last few years is monitoring active, established players who are between 35 to 45 years old, and checking the change in average rating for this group over time. We chose this group, in particular, because we expect, on average, that such players do not have strengths that change appreciably. From about the mid 1990s to 2001, the average rating for the 35- to 45-year-olds dropped between 20 to 25 points per year. Since 2001, the average ratings have been increasing 10 to 15 points per year. Our goal, though somewhat arbitrary, is to increase ratings back to the level they were in 1997. We're about 40 points below this level, so it will still be a few more years until we're at the right place.
I should mention, as an aside, that while we're trying to maintain the level of ratings for this age group, it doesn't mean that we think that a 1600-rated player in

the 35 to 45 age group, in the year 1996, has the same playing strength as a 1600-player in the year 2006. It is possible that chess knowledge has improved in 10 years, so that everyone has simultaneously gotten better. There is no way for the rating system alone to track simultaneous improvement in playing strength because ratings are only relative measures of strength, not a measure of strength on an absolute scale.

HG: Are rating floors useful for combating deflation?
MG: I feel quite strongly that rating floors are an inappropriate method to combat rating deflation. In fact, I am convinced that floors distort the ability of the rating system to measure playing strength. My understanding is that rating floors initially came about as a way to discourage players from sandbagging. Rating floors eventually became an acceptable part of the rating system because players generally do not like the possibility of their ratings dropping without some sort of lower bound. The reason why rating floors are a poor way to address deflation is that the real source of rating deflation, players improving at a rate faster than the rating system tracks, is not addressed by floors. The gain in rating points to the system from floors comes typically from defeating overrated players at (or near) their floor. It does not make sense to me that the opponents of players at their rating floor should be the ones to gain extra rating points. My feeling is that rating deflation is much more appropriately addressed by feeding rating points to the players that deserve it—the ones that are showing evidence of improvement.
HG: Is there a theoretical basis behind the bonus system?
MG: A part of the bonus mechanism in the USCF rating system is mathematically based, and part of it isn't. When the ratings committee originally calibrated the bonus system, we wanted to give extra rating points to players when their performance in a tournament was so good, it would only have occurred 10% of the time. The determination of what constituted a top 10% performance was the mathematical derivation. Then we adjusted the formula so that, in effect, players with high values of K (corresponding to players with low ratings) didn't need to have total scores much higher than expected scores to receive bonus points. This particular aspect of the bonus formula did not correspond to any mathematical model. For our bonus system, the ordinary gain in rating, K(Wactual – Wexpected), must exceed Bm, where m is the number of games in which a player competes, and B is currently set to 10. The lower the value of B, the easier it is to earn bonus points, and the more ratings are inflated. As mentioned previously, the ratings committee has been monitoring the increase in average ratings, particularly for 35 to 45 year olds. When the rating levels in the rating pool reach a point comparable to 1997, we will recommend increasing the value of B in order to help stabilize the dynamics of ratings.
HG: Can you comment on the relationship between USCF and FIDE ratings? It seems like USCF ratings are no longer grossly inflated compared to FIDE ratings, at least for lower rated players, as they once were.
MG: For a long time, it seemed to be of great interest among USCF administrators to align the FIDE and USCF rating scales. After all, both rating scales are based on the Elo system, so they logically should be producing comparable ratings. But, for some reason, the scales are not comparable-USCF ratings, as far back as anyone can remember, have been higher than corresponding FIDE ratings by as much as 100 points. I, personally, am not aware of the reasons for this initial discrepancy in FIDE and USCF ratings.

My attitude, however, is that the USCF has really no reason to try to align its scale to that of FIDE, or to any other system. The FIDE rating system is by no means a gold standard, so I do not believe we should be judging the validity of the USCF rating system by how well it corresponds to the FIDE system. I think a much more reasonable goal for USCF ratings is for the system to be self-consistent, which is a difficult enough problem. I think that there are several reasons for the USCF scale to operate independently of the FIDE scale. First, the two organizations use very different sets of rating formulas; the FIDE rating system more closely follows Elo's original formulas, whereas the USCF system's formulas have evolved into a much more complex system. Secondly, the two sets of formulas cater to different pools of players; the USCF formulas recognize the large number of scholastic USCF members who are quickly improving, whereas the FIDE formulas do not. Thirdly, FIDE events tend to be tournaments with slow time controls, whereas USCF events (rated under the regular rating system) can be as quick as G/30, so the nature of the tournaments can differ quite a bit.

I think it is of practical interest to know how FIDE ratings correspond to USCF ratings because we need to estimate USCF ratings for foreign FIDE players who have not yet competed in USCF tournaments. The conversion from FIDE ratings to USCF ratings is a task that the USCF Ratings Committee carries out every couple of years. To do so, we obtain a list of players who have both (established) USCF and FIDE ratings, and who have been active players in several prior years. We then examine the relationship between USCF and FIDE ratings in order to determine a conversion from the FIDE scale to the USCF scale using appropriate statistical methods. One of the most striking results of this type of analysis is not so much what the conversion formula turns out to be, but the amount of variation that exists in players' USCF ratings for a fixed FIDE rating. For example, for players who are currently rated close to 2200 FIDE, their corresponding USCF ratings range from less than 1800 to over 2450. One of the concepts that this FIDE conversion analysis illustrates is the amount of uncertainty connected with ratings. Even if a 2200 FIDE-rated player, on average, has a USCF 2200 rating, the enormous range of USCF ratings around 2200 highlights the idea that players' ratings are inexact estimates of true underlying playing strength.

HG: The USCF currently has two rating systems for over-the-board chess: standard ratings and quick chess ratings. Is this justified?
MG: Yes. Quick chess requires some different skills than slow chess, such as the ability to choose good moves in complex positions, without the luxury of calculating different lines of play. Some strong players do not have this talent, which is why they probably would have quick chess ratings that are comparatively lower than their regular ratings.

HG: 200 rating points is significant for calculating rating floors, for sectioning tournaments, and in the design of the old title system. Is there any mathematical significance to the number 200 in the current USCF rating system?
MG: My understanding is that giving special significance to the number 200 is arbitrary, and certainly has no mathematical importance. By analogy, the scale of Elo ratings generally ranging from 100 to roughly 2800 is arbitrary. It is easy to imagine that if the rating scale were set up to be a different range of values, then some other round number other than 200 would have arbitrary meaning.

HG: Chess players sometimes boast that if they could manage to play only lower rated players then they would have an easier time gaining rating points. It seems like the opposite is actually true.
MG: Yes, as I mentioned earlier, the winning expectancy formula is overly optimistic for the higher-rated player, which means that, on average, higherrated players will lose rating points if they primarily tend to play lower-rated players. I believe that the main reason for the inaccuracy of the winning expectancy formula can be traced back to the imprecision of published ratings. It can be shown, mathematically, that if the winning expectancy formula is applied to published ratings, but the published ratings are not exact (and that the true rating parameters are equally likely to be above or below the published rating), then the true expected outcome of a game is a value that is between the winning expectancy computed on the published ratings and 0.5. For example, if according to published ratings a higher-rated player's winning expectancy against a lower-rated player is 0.8, then if the two players compete in many games, the actual average result will tend to be between 0.8 and 0.5 (i.e., lower than 0.8).

In a little bit of self-promotion, this problem with the winning expectancy formula is easily corrected in the Glicko and Glicko-2 systems because the Glicko formulas explicitly account for the uncertainty in ratings when calculating expected scores.

HG: What do you see as the biggest challenge maintaining a working USCF rating system?
MG: Way too many players, including people in administrative positions within the USCF, incorrectly view ratings as measures of achievement, or worse, as measures of self-worth. The fact is that ratings are solely intended to measure playing strength. Any use of ratings needs to be consistent with their predictive purpose if the rating system is to work properly. The incorrect way of interpreting ratings leads administrators to favor concepts like rating floors (which, as I mentioned earlier, threaten rating accuracy), and overly simplified rating formulas. Also, the way tournaments are organized creates interesting incentives to manipulate one's rating. Because many tournaments are sectioned and prizes are awarded based on rating ranges (e.g., under-2000 tournament sections), some players are tempted to lower their ratings by purposely losing, allowing them to be competitive for lower-class prizes. The solution to many of these problems, I believe, is to bring back a properly implemented USCF title system, and have the focus of player achievement shift away from ratings and move to titles. Once you earn a title, it is truly an achievement because it is never taken away. In my version of an ideal tournament chess world, not only would people care more about titles than ratings, but also tournament culture would revolve around titles. Tournaments would be sectioned according to titles instead of ratings, and prizes would be awarded to players within title groups. Technically, I believe this would work because as a player improves and has better results, he will accumulate higher-level titles, and would therefore have to compete in higher title tournaments or sections in order to win prizes. It is important to remember, though, that an accurate rating system would need to underlie the title system, because it is the rating system that would make the title system meaningful. The key aspect to all of this is that the USCF would really need to put in the effort to make titles seem worthwhile. It is a very difficult task to change the current mentality surrounding the meaning of ratings, but I think with some effort it is possible.

HG: Thank you, Mark, for your time and your expertise.
MG: Thank you.

Further Reading: Glickman, Mark E. (1995), "Chess Rating Systems," American Chess Journal, 3, 59-102. .