The Performance Rating Algorithm (PRA) and the Tournament App

by Tony Berard

Prolegomenon

This introduction will touch on some key features of my rating system. After reading this introduction, the reader should hopefully feel that this rating system needs to be adopted by the many many places in the world of competition where it would render far better service than what they currently have—the Elo rating system. The Elo rating system is generally acknowledged as being correct. So, any challenge to it can really only be in the area of speed. A faster rating system that generates the same set of ratings would be adopted if we could demonstrate such an increase in speed. I have such a demonstration—read my case study on the 2013 US Open and its follow up. I discovered how to get the most power out of my rating system because of this tournament. Once I operated my rating system at full power on that one tournament's worth of data, it produced a set of ratings for the 24 players in that tournament that the official system took 19 months to approach. I will return to discussing this tournament, but I need to explain the components in the rating system first.

The rating system has two major parts—the PRA, which is the performance rating algorithm, and the Tournament App, which analyzes tournaments or matches.

The PRA is an acronym for Performance Rating Algorithm—the rating system I developed. It is composed of two parts: The Basic System and the Boosting System. Each of these systems will be described in their entirety so that a programmer can code the algorithm into the software that a company or organization uses for rating information for its population of players. The PRA is an automatic system that updates a player's rating upon completion of each game, much like all the other rating systems out there today, which includes the Elo system. I will discuss the rating system in its entirety as the paper moves along. Right now, I am just introducing the parts with a brief discussion about them.

The Basic System is an accurate and reliable rating system on its own, but it is slow and subject to rating inflation and deflation problems like the Elo rating system is currently demonstrating. The Basic System will not change the total points in the system, but people leaving the population will cause the average rating to slide up or down.

The Boosting System has been developed to speed up the process of getting the population in spec (within 100 points of each player's true rating). The Boosting System watches the Basic System and will at certain times boost or crash a player's rating if the conditions warrant it. Early in the boosting process, the boosts and crashes can be quite dramatic. However, these abrupt movements are not wreckless as they have been calibrated through a simulation to be just the amounts needed to bring as many players into spec as quickly as possible. The Boosting System has a variable that is global in scope that allows monitoring of the whole population to check to make sure the average rating is not deviating from 1500. If it goes up, we subract the amount needed from the average rating to get it back down to 1500. If it goes down, then we add to get the average back up to 1500. The Boosting System needs to be damped down at some point, and later it needs to be turned off. Once the bulk of the population is in spec, the Boosting System is no longer needed. However, this global variable that keeps the population average at 1500 needs to be kept intact.

The Tournament App is a software application I developed that allows a qualified person called a Ratings Administrator to process a tournament. Among the tools the app has is a statistical test called The Ratings Goodness of Fit Test. If a tournament fails this test, then it is incumbent upon the

Ratings Administrator to discover (if possible) why the tournament failed. The Ratings Administrator has the power to change initial ratings, among other things, to get at the truth of how the ratings should be both before the tournament and after the tournament. The integrity of the ratings is the sole motivating factor of the actions of a ratings administrator. If fraud is discovered by the ratings administrator, then bad stuff will happen to the perpetrator (of course). As much as possible, the ratings of the rest of the players should be ascertained and updated. The Tournament App provides a lot of information, and it is probably best to let it be automatic. However, the Ratings Administrator's sole task is to reconcile the statistical information with the real world information to get it right. So, we shall allow a Ratings Administrator to override the recommendations of the App if a good reason exists.

We want the correct people to win the tournaments and titles and title shots. Quite a number of times, I have found that with the current tie-breaking systems in place, these benefits have gone to the wrong people. The Tournament App will be fully described as well for a programmer to code.

If just the Basic and Boosting Systems are used from my rating system, that would be quite a bit better for an organization than the Elo system. However, the benefits of the Tournament App are enormous and will be explained next.

The Tournament App does several things, and it uses some other components of the rating system to accomplish them. The first thing it does is use the game result data to measure the performance of each player with the P-Zero scores. Using regression, we regress the initial ratings on these P-Zero scores to get a regression line. This regression line is the Performance Rating line.

We can get a couple of things from here: the linear correlation coefficient (r), and the Ratings Goodness of Fit Test. The linear correlation coefficient is a sample statistic, and we can do a hypothesis test on it to see if the players are expected to finish in the order of their ratings ranked from first to last or not. If this fails, then it's anybody's tournament!! If it passes, then if the highest rated player fails to win it, then we can consider it a choke of sorts. The Ratings Goodness of Fit Test checks to see if the players perform at their rating levels. The performance ratings need to be in spec with the initial ratings at least within a certain level of significance.

If this test fails, then we are obligated to find out why. For example, the 2013 Cap D'Agle tournament, which was won by an elderly Karpov, with a rating below 2700 now, finished ahead of two 2700+ guys, failed this test. At first I thought that games were being thrown to Karpov, but this hypothesis did not stand up to scrutiny. I tried giving Karpov a 2700+ rating, but this didn't solve the tournament either by getting it to pass because it still failed. I kept searching and eventually discovered that they were changing the time controls as the tournament went along. That sort of thing would cause the test to fail because a player's rating is based on a certain set of conditions being met. A player will play at his or her rating under these conditions. If you change these conditions, then the player will play at a different rating level. This is why players nowadays have three ratings: classical, rapid, and blitz. Some players perfom better against the population under a faster time control, but some do worse. I was satisfied that fraud did not take place during the 2013 Cap D'Agle tournament because the amount of out of spec appeared to be explained by the changing time controls, and so Karpov pulled a G-score of 207, so he's still got it!! The G-score will be explained soon. Also, regarding this tournament, no one will tell you that anything appeared to be amiss, but my rating system failed this tournament. So, there was a problem with the initial ratings not fitting with the performances of the players, but upon closer review, I discovered that the reason why was not due to fraud. So, in the end, everything was ok.

Knock out style tournaments will also cause a tournament to fail because higher rated players can get knocked out in the early rounds. Barring anything like that or the discovery of fraud, the test is

telling us that the people who are out of spec need to be adjusted to their performance ratings and run the tournament again with these replaced initial ratings. I did this for the 2013 US Open, and found it was more than a year and half faster than Elo in getting to the correct set of ratings for the players in this tournament. The official TPR's are not good enough to detect these three players (Kamsky, Sevian, and Troff) as being the ones who did not perform as their ratings said they should. My performance ratings using regression with the initial ratings on the P-Zero scores is good enough to detect these players performing differently than their ratings. And, it may be good enough to do better than this. The 2015 tournament for the US Championship just completed, and I found Wesley So to be out of spec by 107 points. However, four players, including Sevian and Troff, were in spec but with a residual magnitude larger than 75. I am calling this rather large residual that is still in spec an upward pressure or a downward pressure on a rating. Over the next few months, we can see if Sevian, Troff, and Holt increase in rating because they had upward pressures; and we can see if So and Naroditsky go down in rating because So was out of spec to low and Naroditsky had a downward pressure on his rating. I will monitor the ratings of the 12 players in this tournament to see if these players change in the directions I have indicated while the rest of the field remains relatively static.

If we have an old tournament with missing initial ratings data (see the Syracuse 1934 Tournament Case Study for more information) or new players to the system who are in need of a tournament ratings calibration, we do a similar thing to the performance ratings. We match up the initial ratings that we do have with the P-Zero scores for those players. This time, the regression line is called the Initial Ratings Estimator Line. We use the Ratings Goodness of Fit Test to see if the line is good. If it fails, we omit the largest out of spec, and run it again to see if we have a good Initial Ratings Estimator Line this time by checking it again with the Ratings Goodness of Fit Test. If it passes, we use the Initial Ratings Estimator Line to get the rest of the ratings. Then, we rerun the tournament again with a full set of initial ratings. So, the regression can give us performance ratings or initial ratings, and these are both accurate because of the Ratings Law of Large Numbers.

This law is the underpinnings of the Tournament App. Players perform at their rating strength. So, we can tell if the tournament went as it should because the P-Zero scores are a very accurate measurement of their performances during the tournament. The accuracy of the P-Zero scores are assured because the simulation I wrote that found the PRA. So, the correctness is assured at every turn.

Thus, when Sam Sevian became an actual grandmaster (the youngest in US history), I was not even surprised because my rating system measured him as one at the 2013 US Open, which was quite a while before the official system did. I have a graph showing how each month when the new Elo ratings come out for these 24 players, I test to see how well they fit the ratings my system developed for them. I can clearly show that my system was 19 months ahead of the Elo system, but Elo used the many tournaments worth of data those 24 players participated in during that time interval of 19 months. My rating system used only that one tournament's data..My rating system saw that two boys (yes, there were two boys in this tournament unbeknownst to me) were both above 2500 in their performances and an elite player named Kamsky who was rated too high. It saw that the rest of the ratings did not need adjusting. All of this was discovered slowly over the next 19 months as the Elo ratings for these players converged to the ratings my rating system had as measured by the test statistic as a time series from my Ratings Goodness of Fit Test.

The main reason why my system did so well was because it identified three players who were out of spec (more than 100 points away from the player's true rating). My system has provisions within it to tell me which direction and by how much to adjust each player out of spec (i.e move the initial rating to the regression line). Once these three players were fixed (to the nearest whole number, not just the nearest 25 or nearest 50), then the rating system was allowed to process the whole

tournament, and it produced the set of ratings that were far ahead of what Elo would only discover many months later with a lot more information. So, this is my absolute biggest piece of news about my rating system. I initially left Kamsky's rating alone because he was an elite player. However, over time, his rating dropped. I learned, to my surprise, that what the rating system is detecting needs to be respected. In hindsight, it is preposterous to say that the two boys need to put at grandmaster level, and the elite player needs to be dropped by over a hundred points and leave the rest of them alone (and some were literally on the border of being a grandmaster). Yet, time has shown that that was exactly what the situation was, and only my rating system could "see" that. My second biggest piece of news about my rating system (the G-score) will be presented next.

The last component of my rating system is the G-score and tournament strength computations. The G-score is a measure I devised to ascertain elite performance in a tournament. It is resistant to the problem of ratings inflation/deflation, so it allows comparison of players across the generations. It also has a stamina factor built into it, so we can now compare performances from tournaments of different lengths. Such a measurement is an advance in our methods to discuss the great tournament and match performances. I can even use ratings information from other rating systems, and I still get a valid G-score because my rating system checks these initial ratings provided by the other rating system against the performances that these players put up during the tournament. The regression will scale the P-Zero scores to fit in with that other rating system's ratings. This is why I can use other rating systems to get G-scores. The G stands for greatness, and I call the formula my greatness formula. With this formula, I have found that Karpov's 1994 Linares performance was not the greatest in chess history, although it was very good.

Recently, Fabiano Caruana had a fabulous performance at the Sinquefield Cup 2014 that is touted as being the greatest of all time over even Karpov's great tournament. I have discovered a number of tournaments better than Karpov's 1994 result, but the greatest performance that I have discovered thus far is Tarrasch's Vienna 1898 performance where he scored a G = 390. This performance might not ever be beaten because the number two performance of all time was Pillsbury's performance at this same tournament (Vienna 1898) in which he scored a G = 374. It was a long tournament with Tarrasch and Pillsbury battling it out all the way. In the end, they tied, so they held a 4 game playoff. Tarrasch won it by a score of 2.5 to 1.5. So, the long tournament (with a large stamina coefficient) combined with neither player letting up at the last round or two with draws (or even losses as in Zuckertort's London 1883 where he "let up" because no one could catch him with three rounds to go even if he had 3 losses, which he did (because he started taking opium (he was a long time user by the way, and the other players knew it for quite some time))). Had Zuckertort not let up, he could have had it, but instead, he gets a G = 368.43, which is only a few points behind Pillsbury's Vienna 1898 great performance. For the record, here are the top two conventional wisdom's greatest, and the top two of what I have found with the G-score to be the greatest.

Conventional Wisdom:
1) Caruana 2014 Sinquefield Cup: G = 302
2) Karpov 1994 Linares: G = 308

My G-score:
1) Tarrasch's 1898 Vienna: G = 390
2) Pillsbury 1898 Vienna: G = 371

A bit of observation has enabled me to demarcate the G-scores into useful categories. Anything over 100 makes the list, but a G-score of 100 is not dominance of the first water. The following chart is what I use when discussing G-scores.

100 – 149.99	Making the G-List
150 – 199.99	Very Good!!
200 – 249.99	A Dominating Performance
250 – 299.99	A Threat to the World Championship
300 – 349.99	A Generationally Great Performance
350 – 399.99	A Historically Great Performance
400+	Beyond Human Comprehension

Thus far I have not found a 400+ performance in either a tournament or match. Steinitz was the closest. If the match continued and Steinitz kept winning, he certainly would have exceeded 400. But, that did not happen, so even Steinitz has not exceeded 400.

An integral part of the G-score formula is the resistance coefficient. This measures the amount of resistance a player offers on an absolute scale. I can use this component from the formula from all of the players in a tournament weighted by the number of games each plays in the tournament to arrive at an overall strength of the tournament that is simply the best measure for this sort of thing yet devised. I have constructed the definition that the number 2 player in the world has a resistance coefficient of 1.0. The number 1 player will have a resistance coefficient, r, greater than 1.0. Everyone else will have an r less than 1.0 but greater than 0.0. If we multiply r by 100, which is what I do, then we get the tournament strength expressed as a percent. It isn't really a percent, but it kinda works. Matches with the #1 and #2 players will have a tournament strength greater than 100 percent. It depends on the size of the gap between #1 and #2. Bobby Fischer had opened up a 125 point gap in rating between him and Boris Spassky, so the match between these two players would be the strongest "tournament" in history at 102.2321 percent. To compare, another #1 versus #2 was Steinitz versus Blackburne in 1876. That one had a tournament strength of 101.4643 percent.

So, the G-score is a very valuable addition to our arsenal of weapons with which we assess the great performances. It splits all the hairs needed to rank all of these great performances. I have found a number of them above Karpov's. I have compiled a list of these G-scores called my Greatness List. I hope to answer definitively who turned in the best performance of all time. When I cite examples of tournaments with measurments from my rating system, it is because I have put that tournament in the archive so that is can be inspected.

So far it is Tarrasch's performance, for a tournament performance, that is the greatest; but Steinitz's performance against Blackburne is the greatest chess performance of all time. Steinitz and Blackburne were the top two players in the world in 1876, according to Chessmetrics. Steinitz defeated Blackburne by the overwhelming score of 7-0. Yes, Bobby Fischer wiped out Mark Tiamanov 6-0 and Bent Larsen 6-0 in their individual candidates matches leading to Fischer taking the title away from Boris Spassky in 1972. But, Tiamanov and Larsen were not the second best player in the world. Spassky was the number two at the time, and Fischer beat him soundly, but not a wipe out like what Steinitz did to Blackburne. So, Steinitz's performance against Blackburne is the greatest in chess history, and he stands at the top of my list of G-scores for it at G = 397.44.

The strongest tournament conversation comes up from time to time. There have been a number of schemes developed to say that this tournament or that tournament was stronger than this other tournament. The G-score formula has a resistance coefficient in it that measures the amount of resistance that the opponent offers. This component can be calculated for each player in the tournament and weighted by the number of games played by that player. Thus, the weighted average of the resistances of the players would be an inflation proof measurement of the strength of a tournament. I have compiled such a list of these tournaments along with the greatest tournament performances of all time. It usually is the case that a great performance occurs at a elite tournament or international tournament, but not necessarily a super grandmaster one. If the tournament is a

supergrandmaster tournament, then they usually beat each other up too much for a player to dominate enough to get a good G-score.

I will eventually finish the work of combing through chess history, and I want to go through Gino Di Felice 's massive database of games and tournaments. Maybe I might even get to collaborate with him to do this work. He has the biggest pile of data, and I have the best data processing system for this kind of data. So, it would benefit the chess world if he and I were to collaborate.

Who stands to benefit from my rating system? There are many organizations that would benefit. Chess is the most obvious, and there are millions of chess players all over the world. There are chess tournaments going on all the time all over the world. With my rating system, we could find out the true strength of any up and comers immediately. My rating system allows immediate calibration and correction of ratings. So, we can now know that in some XYZ tournament that player such and so is a viable candidate for the pool of players used to find the next challenger to the world championship title. As it is now, that player goes unnoticed by the chess world at large unless that player does well in a number of other tournaments to build up a large enough rating because the Elo system is very slow (but accurate).

The games or competitive events of pool, scrabble, checkers, backgammon, shogi, go, Chinese chess, abalone, othello, many video games, ping-pong, Arimaa, fencing, and racket ball among others would also benefit from my rating system. If I were to get a few of these games to adopt my rating system for its population of players, then it is conceivable that my rating system would get used a billion times each day. The sheer magnitude of immediate applicability for my rating system cannot be overstated.

I developed the PRA, which stands for Performance Rating Algorithm, for the eventual population of Tines and Barbs players. But, as I already stated, the algorithm has wide applicability for many other things. This algorithm rather quickly whips a population into shape from what I call the initial state of chaos, which is every player set to a rating of 1500. It certainly does it far faster than the Elo rating system. My rating system is capable of processing tournaments in chess's history as well. I have used ratings data from older tournaments that had errors or missing information. My rating system can fix these problems and give a complete and accurate picture of what happened if enough information is present even though some of that information may not be correct.

One example is the Syracuse 1934 tournament where a young Samuel Reshevsky won the tournament ahead of the number two player in the world at the time, Isaac Kashdan (according to Chessmetrics). But, Chessmetrics has ratings data on only 7 of the 15 players, and Reshevsky is not one of them. This tournament is one of my case studies, and I completely show how my rating system fills in all of the rest of the picture for this tournament, including the amazing G-score for Reshevsky of 270.26.

There are quite a few tournaments on the Chessmetrics website that have missing information like this. The worst one I have found was the 1957 World Junior Chess Championship. The tournament was won by William Lombardy, and he achieved a perfect score of 11-0. Chessmetrics had only one rating for this tournament—William Lombardy's. The performance ratings for the remaining 11 players were all identical because it was based on one game against the rated William Lombardy, which were all loses of course. So, I used this performance rating for the highest scoring player to Lombardy. I found a couple more ratings somewhat removed in time from 1957 for a couple other players on the Chessmetrics site. These I used to build an initial rating formula for the rest of the players with regression. With that extremely limited information, I found that Lombardy scored a G = 234.89, which is a dominating performance.

One tournament I processed was Sofia Polgar's amazing performance in Rome 1989. She blew away the field with an amazing 8.5 out of 9. I can only find the ratings of Sofia and six of her strongest opponents. The first two rounds, she said she played against lower rated players than herself. She was over 2300, so I gave these two players a 2300 rating. The rest of the ratings are higher than Sofia's rating by quite a bit. She said (on her website) that she had never performed as well before or since that tournament. I processed her nine games, which is all the information I had, along with the ratings as I stated. Sofia scored a G = 255.15 in the Rome 1989 chess tournament, which is at the level of a contender for the world championship! I have no way of checking this against the P-Zero scores to see if it passes the Ratings Goodness of Fit Test. But, the tournament is well documented to have occurred, so I have no reason to doubt its veracity. The 14 year old Sofia reached for the stars that tournament. The highest G-scores of her much stronger sister, Judit, by comparison are the following:

Isle_of_Lewis_1995: G = 203.15
Madrid_1994: G = 203.61

So, who was stronger between Judit and Sofia? Anyone you ask will instantly and unequivocally say Judit, and they might even add the words *by far*. I will tell you, though, that for one tournament, Sophia surpassed anything that Judit ever did.

I have done a number of tournaments by Garry Kasparov, and he has never had a G-score above Karpov's 1994 Linares tournament, which he scored G = 308.89. The two closest times Kasparov came (as of this writing) is Tilburg 1989 when he scored G = 302.54 and Batumi 2001 when he got G = 305.55. Both Kasparov and Karpov have many many tournaments I have not processed yet. I have processed the main ones others have said were their best. But, Tilburg 1989 was supposed to have been Kasparov's best, but I found that he did better in Batumi 2001, which was still a bit shy of Karpov's best. But, I am always working on this stuff, so I may find other diamonds in the rough, so to speak, by these two greats. As it stands now, though, neither Karpov nor Kasparov have sufficient G-scores to challenge Tarrasch at G = 390 (Vienna 1898) for the greatest tournament performance of all time.

The G-score actually compares speeds of rating gain with the two modifiers of resistance encountered and amount of stamina required. In physics, we learn that we can compare the speeds of two moving objects with meaningful results. The G-score formula goes beyond this by having two multipliers—resistance and stamina. The resistance coefficient is a massaged number to get it to be inflation proof as well as being able to be ported to other rating systems. The essay I wrote on the greatness formula is included in this document as a chapter and explains this massaging I am talking about. The stamina coefficient starts at 1.0 and increments by 0.05 for each successive game played. Thus, there is a balance struck between the resistance and stamina. Longer tournaments provide a larger stamina coefficient, but players are unable to keep up an unbeaten streak. Tarrasch's Vienna 1898 was a very long tournament, and he scored a little above 75 percent. But, the resistance he encountered was 0.89 with a stamina coefficient of 3.05. The two coefficients multiplied together are 0.89 * 3.05 = 2.7145. By comparison, Pillsbury had 0.88 * 3.05 = 2.684. Had he won the playoff, his G-score would have been higher than Tarrasch's G-score.

For the last part of the introduction, I will discuss what Dr. Lasker (the world champion for 28 years) attempted with chess. He is known in the chess world as a great champion with the greatest longevity. However, it is less well known that his Ph.D.was in mathematics. He wrote a book about chess called *Lasker's Manual of Chess*. In it, and I have heard no one else mention this, he set out to prove chess mathematically, which would make the outcome from any initial position a certainty. We would simply inspect the postion and determine which chess theorem in his maual applies, and then we could conclude it was a win for White or a draw etc.

Let us contrast this with a random event with a certain probability assigned to it. Based on the ratings of the players, we will say that White has a 40% chance of winning the game from a position of equality if White is about 150 points lower in rating than Black. This is completely divorced from the contents of the chess position! Yet, this is what has happened over time to the evaluation of chess outcomes. However, this assessment generally occurs before the game begins. Once the game is underway, then we look to the position to see who is winning or has an advantage etc.

So, the ideal rating system (i.e. the PRA and Tournament App) gives us a good method to help us determine who will win a particular game, and not Lasker's position proof method. Such a rating system also tells us who the strongest players are in terms of their G-scores. I found a statement that Bobby Fischer said regarding the 1948 tournament that Botvinnik won. He didn't really win the tournament as top Russian GM's were ordered to lose enough games so that Botvinnik would win. He was their guy, and he stayed their guy for many years. Bobby Fischer said that Reshevsky was the strongest player in the world at the time and would easily have defeated Botvinnik in a match for the title. The Russians drew each other and fought hard against Reshevsky, which made it unfair to him. The same kind of thing happened to Bobby Fischer later on, but he made a big stink about it. That is why the lead up to the title in 1972 was matches against the other candidates. We all know Bobby Fischer blew out Tiamanov and Larsen 6-0 each. But, Reshevsky didn't make a stink about what transpired, so he didn't benefit from match play like Bobby Fischer did. So, we can check Bobby's assessment with the G-scores from that year, and that is a future project of mine. I will seek out all the tournaments I can find from 1946, 1947, and 1948 that involves the top players of that time and see who garnered the highest G-scores. Similarly, the G-scores can be used to determine the best players from any time period or for any competitive modality. I am very excited to be embarking on this quest for games like Scrabble and Backgammon, and if I could do the ratings for something like fencing, that would be amazing. The same rating system works for all of these modalities because there is sufficient complexity in them for the practitioners to be able only to be proficient in them to varying degrees, but not a complete master. However, for now I will just work with chess: its history and its current events.

As I get tournaments done, I will put them in *The Greatness of Chess Archive* at the address https://docs.google.com/file/d/0By2fC4qQL9sKNjdrSkVLZHN4WUU. This is my major document that chronicles chess performance. It is the work of a scientist applying the objective method. I supply my data (the ratings of the players and the games played). I run the data through my app to get the conclusions reached. I have had to invent and modify some of the standard fare of statistics, but all of the things I have done to get my rating system to work are all chronicled in this paper..

An Outline of the Project

Here is an outline of the Performance Rating Algorithm and Tournament App and their parts that will be described in this paper (or possibly future other papers).

I) The entire rating system including the parts

A) The Basic System

B) The Boosting System

C) Note: The Basic and Boosting Systems are "automatic" and operate at the individual game level.

D) The Tournament App (automatic with possible manual override) which includes the following parts.

1) The initial rating formula for missing initial ratings via regression.

2) The calculated P-Zero scores

3) The calculated PR scores

4) The Ratings Goodness of Fit Test

a) If the tournament passes, allow the automatic systems to adjust the ratings as normal.

b) If the tournament fails, then do one of the following:

i) Allow the automatic systems to adjust as they normally would, but make an administrator's notation about the tournament as to why this occured. As long as fraud was not why the tournament failed, then everything should be fine.

ii) Adjust the desired out of spec ratings using the PR's (based on the P-Zero scores). Also, it is permissible to adjust the ratings with upward or downward pressure (ratings not out of spec, but having a residual magnitude from 75 to 99).

iii) Rebuild the initial ratings list by deleting the out of spec ratings or ratings having pressure, and create an initial ratings formula to rebuild the deleted initial ratings. Any decisions the ratings administrator makes needs to be documented.

E) Note: The Tournament App is automatic, but when the tournament fails the Ratings Goodness of Fit Test, the Ratings Administrator can override the system if there is a good reason to do so.

F) Note: Linear regression is justified when adjusting or rebuilding the initial ratings by the Ratings Law of Large Numbers--a law I discovered about ratings. I am the first person to see the law working, and I will explain why it works. It is what gives any large adjustments their power of correctness in a single adjustment.

G) Upon passage of the Ratings Goodness of Fit Test or allowing the automatic adjustment to kick in, compute both the G-scores and the Tournament Strength.

II) The next paper is a bunch of discoveries about chess history that were brought to light or overturned by my rating system. An important one is that Gary Kasparov is not the greatest chess player ever. Quite a number of chess players have G- scores in excess of anything that Gary has. Another recent one is that Fabiano Caruana's 2014 Sinquefield Cup performance was not the greatest ever, exceeding Karpov's 1994 Linares performance. The greatest performance I have discovered so far is Tarrasch's 1898 Vienna performance in which he obtained a G-score of 390. Karpov's 1994 Linares performance was G = 308. Caruana's 2014 Sinquefield Cup performance was G = 302. Had Caruana won the final two games instead of drawing them, he would have had G = 391, which would have given him the greatest performance of all time.

III) One paper will be the source code of the simulation I constructed that enabled me to find the PRA. It will also contain detailed comments on its many parts.

IV) One paper will be the source code of the tournament app I constructed that enables me to ferret out the many truths I discovered about chess history. I need to upgrade it to include the probability of a sweep for the favorite player, a hypothesis test on r, the linear correlation coefficient (if it passes, then the hypothesis of there is a correlation means that the top three rated players should have finished in the order of first, second, and third, respectively; however, theoretically, they should finish exactly in the order of their initial rankings), and the ability of the app to use the ideas in the Syracuse 1934 paper to provide initial ratings for those missing or way out of spec. I also want to have it read in the Greatness List and List of Tournaments to update it after processing the current tournament. I will have other bells and whistles as well.

Part 1: The Performance Rating Algorithm

The Basic System

Each new player that comes into the system brings with him or her the average rating points (Rbar = 1500.0) with them. Thus, we can know immediately how many points P the system should have by multiplying the average number of points per person by the number of people in the system or N. Thus, suppose we have N = 20,000 people in our system. Then,

$$P = Rbar * N = 1500.0 \times 20{,}000 = 30{,}000{,}000 \text{ points.} \quad (1)$$

The Basic System uses the following formula:

$$R' = R + 9.0(S - E) + adj. \quad (2)$$

In words, this is the new rating is equal to the old rating plus nine times the difference of the player's actual score and his expectation. We add to this the adjustment, which is the opposite of the surplus or deficit in the total required points for the system divided by N, the number of players in the sysem.

Using the above example with 20,000 players, suppose we add up all of the points in the system and find it to be 29,875,241. We would say the system has undergone some deflation, but we can fix it. We find that the deficit is 30,000,000 – 29,875,241 = 124,759 points. If we now divide the deficit by N = 20,000, we can add this amount to each player's rating in the system to restore it. Thus, adj = (deficit or surplus)/N = 6.24 points per player, which is the amount to adjust the players by to restore the center of the population to Rbar = 1500.0.

Use of the adj in a rating system is how to prevent rating inflation and deflation. Solving the rating inflation/deflation problem alone makes this system a better system than the Elo system because Elo is now showing very clear signs of inflation. Bobby Fischer's rating at its peak was 2785. In the late 1970's, only one player was over 2700 (Karpov). Today in 2014, there are many many players in the 2700 club with even a few in the 2800 club. Magnus Carlsen has even gotten as high as 2882 with postings on the internet asking for predictions when he will break the 2900 barrier. Could Carslen have defeated Bobby Fischer with the expectation that goes along with a nearly 200 point difference in rating? Of course not. This is the impression generated with these inflated ratings of today, however. So, having this global variable adj is important to avoid all of the issues (and media hype) that goes along with the inflated ratings.

The S is the score that the player received against the opposition. If a player wins, he gets 1.0 points. If he loses, he gets 0.0 points. Other modalities of Tines and Barbs (a board game I invented that has ways of playing it with greater outcome refinement than win, lose, or draw) allow finer subdivisions of the point. Thus, for example, in one particular modality of Tines and Barbs, the point might be split 0.7 and 0.3. We can say that the player receiving the 0.7 portion of the point "won." The player receiving the 0.3 portion of the point "lost." However, the scenario is possible that in one of these non-standard ways of playing this game, that a match might be at 11.7 to 11.3 in favor of the champion. Let us say the champion "loses" the final game 0.4 to 0.6. Well, the champion still wins the 24 game match because 11.7 + 0.4 = 12.1 while the challenger got 11.3 + 0.6 = 11.9. Such an outcome is possible in these other ways of playing Tines and Barbs.

The ratings simulation I devised only uses 1.0 and 0.0 to find the PRA. Draws at 0.5 were not even considered in the simulation. Now, that the PRA is found, the finer subdivisions of splitting the point are simply put into the S spot in the equation. The rating system will work on a game such as

Tines and Barbs by using the S as I have shown. Chess has 1.0, 0.5, and 0.0 for its possible values of S. Just insert one of these into S as appropriate to use the rating system for chess.

The E is the expectation. I have built an expectation function that works very well in the simulations I have run for the rating system. It is the following program fragment in the C programming language listing.

```
double EAComp(double val1, double val2) {
double diffT = 0.0, EA = 0.0;
diffT = val1 –  val2; //This will run from -1800 to 1800 approximately.
if(diffT > 1800.0)  EA = 1.0;
if(diffT <= -1800.0)   EA = 0.0;
if(diffT > 450.0 && diffT <= 1800.0)
      EA = 0.9 + sqrt((double)((diffT - 450.0)/135000.0));
if(diffT > -1800.0 && diffT <= -450.0)
       EA = 0.1 - sqrt((double)((diffT + 450.0)/-135000.0));
if(diffT > 150.0 && diffT <= 450.0)  EA = 0.001*diffT + 0.45;
if(diffT > -450.0 && diffT <= -150.0) EA = 0.001*diffT + 0.55;
if(diffT > 0 && diffT <= 150.0) EA = 0.6 - sqrt((double)((diffT - 150.0)/-15000.0));
if(diffT > -150.0 && diffT <= 0) EA = 0.4 + sqrt((double)((diffT + 150.0)/15000.0));
return EA;
}                                                                                    (3)
```

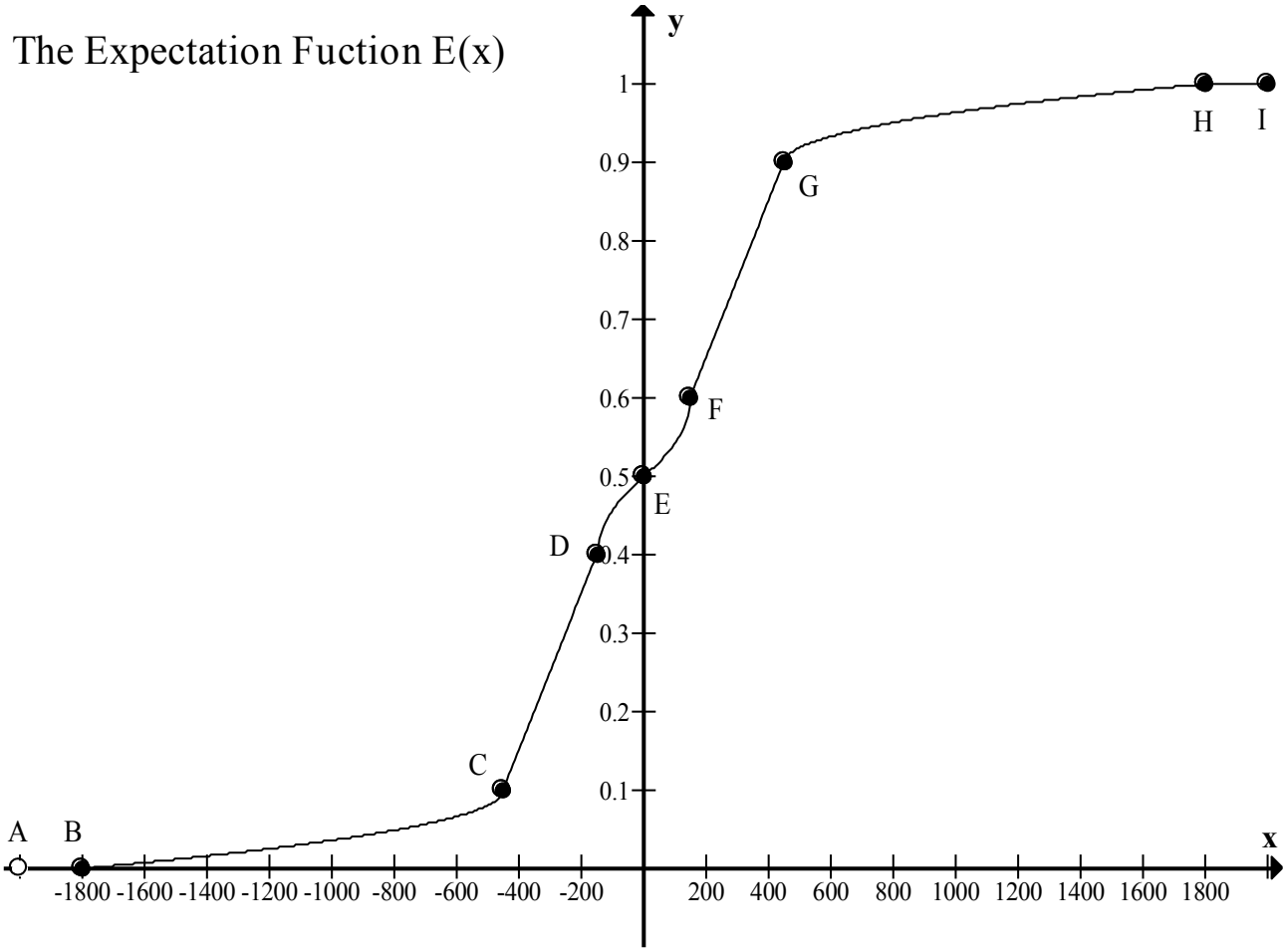
The function EAComp takes two inputs: val1 and val2. This is the rating of the player under consideration and the opponent, respectively. The function uses two variables internally to it: diffT and EA. The diffT is just the difference in the ratings in the order shown in the function. So, if diffT is negative, that means that the player is playing against a higher rated player. If diffT is positive, then that means that the player under consideration outranks his opposition. The variable EA is the expectation of the player. This is the probability of this player's victory against this particular opponent.

Now, diffT will approximately range from -1800 to 1800. If diffT is lower than -1800, then we assign a probability 0.0 for this player's expectation of victory. Alternatively, if diffT is above 1800, then we assign a probability of 1. These are practical probability assignments I have made. If we wait ten thousand years and look at the history of 1800 point differences in rating to see if the lower rated player ever scored a victory, we might find that it had happened. However, empirically speaking, we would never see it in one lifetime. Thus, certainty in an 1800 point difference yielding certainty of victory for the higher rated player and certainty of failure for the lower rated player is justified.

The rest of the function has either lines or parabolic arcs that fit the rules of thumb for rating differences. We expect about a 50-50 chance if two players have ratings within 100 points of each other. The function above has two parabolic arcs from rating differences from -150.0 to 150.0. The EA goes from 0.4 to 0.6. The vertices of the parabolic arcs are at (-150.0, 0.4) and (150.0, 0.6). The "other" point for each is (0, 0.5). Working out the math yields the two square root functions depicted. A graph of the E(x) [or E(diff) or E(Rwhite – Rblack)] function is on the next page.

We shall describe the rest of the expectation function now. For diffT ranging from -450.0 to -150.0, we have the the line y = 0.001x + 0.55. Which in the symbols of the C function is EA = 0.001*diffT + 0.55. The returned EA back to the main program is 0.1 to 0.4. These values are perfect for the rule of thumb of 60-40 chances if the rating difference is between 100 and 200 points. At diffT being -150.0, this is the midpoint of the interval from -200 to -100. The line gives

the EA at 0.4 for this midpoint. It is exactly what the rule of thumb requires. The line is so calibrated to give the perfect result at each of the midpoints from the ranges -300.0 to -200.0, and -400.0 to -300.0, and -500.0 to -400.0. The range of diffT from -1800.0 to -450.0 shows a square root function. The vertex of this function is at Point C (-450.0, 0.1). The other point needed to find the equation of this parabolic arc is at Point B(-1800.0, 0.0). This parabolic arc allows a more rapid descent of the probabilities than a line between these two points. Symmetry will give us the corresponding line segments and parabolic arcs along the rest of the range of values that diffT can assume. Thus, our E is measured in this EAComp function for a player against his many opponents.



The Expectation Function

So, that is all we need for the basic rating system: the player's score (S) against the opposition, the expectation for victory (E), and the adjustment (adj). This rating system is accurate, but it is slow to straighten out the population. It does have a major advantage over Elo in that the adjustment variable keeps the population average anchored at 1500.0. This prevents the rating inflation and deflation problems seen in the ELO system.

A word or two about the problem of rating inflation and deflation is in order at this point. The ELO system exhibits this problem due to different K-factors for the two players under consideration. This causes a different number of points to be added to the winner's rating than what is subtracted from the loser's rating. In isolation, this difference is minute for the whole population. However, when this happens a lot, substantial points can be gained or lost in the system of players that constitutes the population. Too many such points gained is rating inflation, and too many such points lost is rating deflation. Because ELO only looks at two players in isolation, these small imbalances of point changes can result in large scale disturbances in the mean of the population.

Imbalances are also created when players leave the system. If a player started at 1500 and leaves the system at 1700, a vacuum of 200 points is left. If it only happened once, then it's effect would be small. But, that kind of thing happens all the time. That is why the adj variable is so important. Compensating for these imbalances at the end of each day (say at midnight) keeps the population mean anchored at 1500.0, which solves the rating inflation/deflation problem.

The Boosting System

The Boosting System is an advance in the state of the art of rating systems in that we now have a device that inspects the recent performance of the player and decides if a boost up or a crash down is warranted. The decision to boost up or crash down a rating is made at periodic intervals of a player's game history. The size of the boost (or crash) is determined automatically by the boosting system.

What we do is observe a set block of games. For my simulations, I noticed a distinct improvement in accuracy when the block size is 40 games as opposed to 20 games. However, for my website, I want the levels of a player to be differentiated every 20 games, so that at games 100 and 200, we have some kind of recognition with a ceremonious boosting (hopefully, not a dropping). These are possible with block size measurements every 20 games. But, the boosting system breaks down and exhibits instability after game 160. So, a reduced boosting system with smaller coefficients is used from games 160 onward. The reduced system is applied every 40^{th} game. So, at games 200, 240, 280, and 320, the reduced boosting system is run with the smaller coefficients. After game 320, no boosting system is run; it is only the basic rating system.

Break the block size into four groups. So, if you are using block sizes of 20, then each sub-block would be 5 games. We want to capture data points as (game number, rating). In my program, I start the game counter at 0 to match with the zero offset of arrays in the C programming language. Whatever language you are using, you may have this zero offset, or you may not. Just, be aware of this issue because if you have it off, then your data points will consist of a boosted value in one of the blocks of 5 among the other regular values. At game 160 for my population, this error cost me 5% of population to be wrong. I had found appropriate coefficients for this error without realizing it was an error. I got to 85% accuracy at the 160 game mark. When I corrected the zero offset problem, then my points (game number, rating) were accurate, and I had to find new coefficients. When it was thusly fixed, I got to 90% accuracy at the 160^{th} game.

Let's get specific with how to do the boosting system. Each player has a game counter. Mine is the simple variable ctr, but you might use something different. C has the modulus operator %, so I might code it as "if((ctr+1)% ==0)" then do a block of statements related to measuring the block of games. Why do I have the ctr + 1 instead of just the ctr? This put me at game 19 or game 39, etc. This is when the last measurement occurs and the computation of the direction and magnitude of the boost happens.

The variables I used in my program are Rbar1, Rbar2, Rbar3, and Rbar4. Games 0 to 4 comprise five games: 0, 1, 2, 3, and 4. If you are using R as your rating variable, then R(0) is the rating for game 0. R(1) is the rating for game 1, and so on. We would then have that

$$Rbar1 = (R(0) + R(1) + R(2) + R(3) + R(4))/5.0. \quad (4)$$

Repeat this for Rbar2 and Rbar3. Finally, we have the remaining computation

$$Rbar4 = (R(15) + R(16) + R(17) + R(18) + R(19)) / 5.0. \quad (5)$$

We want to know if the player is trending upward or downward. So, I used V1, V2, and V3 in my program to measure the velocity of the player's rating. The three equations are the following:

V1 = Rbar2 – Rbar1 (6)
V2 = Rbar3 – Rbar2 (7)
V3 = Rbar4 – Rbar3 (8)

If V1, V2, and V3 are all positive, the player is trending upward. If V1, V2, and V3 are all negative, then the player is trending downward. Any other combination of positives and negatives for V1, V2, and V3 constitute a player with a stable rating. Players with a stable rating meander above and below the true rating.

I have a variable in my program called direction (or just dir). This variable can be +1, 0, or -1. If it is +1, then it tells my program that this player is trending upward. If dir is 0, then the player is either stable or in process of changing directions. It is important to note, when we boost the rating either upward or downward, it is possible to overshoot the mark. Then, the basic rating system will tend to change the direction of the player's rating as the games progress.

Each player at the beginning should have a direction of 0. If they demonstrate an upward trend, then at the next block assign them a direction of 1. If they demonstrate a downward trend, then at the next block assign them a direction of -1. If they are just meandering, which will be in the majority of cases, then they can stay at a direction of 0. It is important not to change the value of the direction variable from a 1 to a -1 or from a -1 to a 1. This causes instability in the ratings because the players will be in a perpetual state of overshooting their ratings.

Thus, we can have six possible cases of the direction variable on two consecutive blocks: 1, 1; 1, 0; 0, 1; 0, 0; -1, 0; and -1, -1. We will get three coefficients out of these possibilities that ensure a smooth transition of the boosting system correcting the player ratings to get to a stable rating for them.

We need a couple of variables in the boosting formula: peak and trough. As the block of games progress, capture the highest rating and lowest rating that the player achieves during the block. The boost formula is the following:

R' = peak + coeff*(peak – trough) (9)

The crash formula (i.e. to reduce a downward trending player's rating) is the following:

R' = trough - coeff*(peak – trough) (10)

After many simulations, I have found that the three coefficients corresponding to the three directions are 2.625, 0.875, and 0.5. The following code is not C, and it isn't really pseudocode, either. But, it captures the ideas for any programming language, I think. Observe the statement involving the ctr < 20. This is here to change the player's direction at the beginning. If the player gets an upward trend right from the get go, then we can assume that the previous block (which doesn't exist) would have had a direction of 1 for it. This allows the biggest initial jump with the largest coefficient of 2.625 upon the first measuring. The same comments apply for the corresponding statement for a downward trending player.

```
if(V1 > 0 and V2 > 0 and V3 > 0) {
if(ctr < 20) direction = 1
if(direction = 1) R = peak + 2.625*(peak – trough)
```

```
if(direction = 0) { R = peak + 0.875*(peak – trough)
                    direction = 1 }
if(direction = -1) { R = peak + 0.5*(peak – trough)
                     direction = 0 }
}

if(V1 < 0 and V2 < 0 and V3 < 0) {
if(ctr < 20) direction = -1
if(direction = -1) R = trough – 2.625*(peak – trough)
if(direction = 0) { R = trough – 0.875*(peak – trough)
                    direction = -1 }
if(direction = 1) { R = trough – 0.5*(peak – trough)
                     direction = 0 }
}
if(signs of three V's don't all match) direction = 0                    (11)
```

Another comment we can make is for the two-statement blocks. If the direction for the previous block was zero or the opposite direction, we move it one unit (in the manner of counting on a number line) in the direction that the V's say to. This sets us up for the next block of games to get the coefficient correct. Note also that in this code, we have something like "if(direction = 1)". This is asking about the previous block. The three V's being all positive do not automatically mean the direction for this current block is 1. No, it will be 0 or 1. If the previous block was -1, then the current block will be 0. If the previous block was 0, then we assign the current block to be 1. Of course, if the previous block was already 1, then we keep it as 1 for the current block. This is how to get the coefficients right for the boosting.

The boosting system provides its high quality service as outlined here for only up to the 159th game. Game 159 is when the last measurement and boost should be done. Then, the length of the block should double to 40. Most of the players are going to be asymptotic by this time, and to get an improvement we need to measure the rating changes over a longer period of time.

The three coefficients as outlined above should change to omega = 0.15 times their values (i.e. they should shrink). The first measurement under this new scheme should occur at game 199, and boost the ratings accordingly. The boosts and crashes will be smaller now than before. Keep up this reduced boosting system every 40 games until game 319 (or 320 without the zero offset). Thus, a minor boost should occur at games 199 (or 200), 239 (or 240), 279 (or 280), and 319 (or 320). Then, set omega to 0.0 to turn off the boosting system. The regular rating system will take it from there. The idea is explained in the following "code" that isn't really any programming language.

```
if(ctr <= 159) omega = 1.0
if(ctr >159) omega = 0.15 and block size = 40
if(ctr < 319) omega = 0.0
if(V1 > 0 and V2 > 0 and V3 > 0) {
if(ctr < 20) direction = 1
if(direction = 1) R = peak + omega*2.625*(peak – trough)
if(direction = 0) { R = peak + omega*0.875*(peak – trough)
                    direction = 1 }
if(direction = -1) { R = peak + omega*0.5*(peak – trough)
                     direction = 0 }
}

if(V1 < 0 and V2 < 0 and V3 < 0) {
```

```
if(ctr < 20) direction = -1
if(direction = -1) R = trough – omega*2.625*(peak – trough)
if(direction = 0) { R = trough – omega*0.875*(peak – trough)
                    direction = -1 }
if(direction = 1) { R = trough – omega*0.5*(peak – trough)
                    direction = 0 }
}
if(signs of three V's don't all match) direction = 0                    (12)
```

Some Words about the Simulation

The program I wrote to find the PRA (Performance Rating Algorithm) is a simulation of sorts. It is a simulation in that if we could perform the experiment as dictated, then this is what we would get. However, there is something special about this simulation that we could never achieve in real life: knowing the player's true rating. Elo has said that we never really know a player's true rating. But, this simulation runs as though we do have it. This extra piece of knowledge informs the simulation and gives it its power.

It took me quite a while to discover this algorithm, but I wrote a computer simulation to ahieve what would not be possible in real life (knowing the players true ratings, as just mentioned). But, it is possible in the simulation. I had a population of players with fixed ratings. Yes, the program knew the true ratings of these players, but I followed some protocols to enable me to discover what these ratings were. It was kind of like having two people overseeing the simulation: One omniscient of all the details, and one completely ignorant. The omniscient one set up the games and used probability based on the ratings of the players to see who won each game. The ignorant one provided a tentative initial rating for each player and updated it based on who he saw won and lost each game played. As the games rolled along, the ignorant one's rating list for the whole population of players would be checked by the omniscient one to see how it was going along.

I had to devise many rating systems for the ignorant one to use to update his ratings list with. Some rating systems were better than others. Some were just miserable failures. Each one, though, as a concept seemed perfectly reasonable. But, I kept inventing new rating systems for the ignorant one to use. Eventually, I found the PRA. This rating system far surpasses the Elo system in its ability to get a population in spec. Yes, I gave the Elo system to the ignorant one to try. It really didn't do all that well to my great surprise. With a mild disturbance in the initial ratings, it did ok, but not great. With the population set to all 1500's, it didn't do well at all. But, I will get some hard data on the Elo system and other rating systems against this population of 15000 players when I get more time—probably in 2016. Also, my simulation now has an error in it that won't let it compile anymore. So, I have to find and fix that. Along the way, I am going to incorporate more bells and whistles into the simulation to yield more data on how well the various rating systems do.

I created a bunch of rating systems that all failed for one reason or another, except for one—the PRA. The PRA is actually two systems: The Basic System and The Boosting System. I set up a simulation that randomly paired off all of the 15,000 players in my population. Thus, each round consisted of 7500 games. I didn't consider draws—only wins and losses. I could set the simulation to run for however many rounds I wanted. If I ran it for, say, 600 rounds, then that would be 4,500,000 games that the population played. Within the framework of these simulated games, the rating systems I designed could be improved upon, calibrated, or eventually scrapped. All were scrapped except for the PRA which was improved upon with the addition of the boosting system, and it was also calibrated to its ideal state. The two systems (the basic and boosting systems) have a number of coefficients that needed to be found and calibrated.

Now that the coefficients have been found, the system is complete. Well, the rating system is not just this PRA, which includes the Basic System and the Boosting System. The rating system also includes the tournament app, which is fully discussed in the second part of this paper. But, this was just a few words about the simulation. I just wanted to tell you in this section that we have an aspect of omniscience in the simulation that would be absent in real life. So, we get an amazing benefit that an ordinary simulation could not give.

How do we pull this off? I mean to say that we are going to say we know the player's true rating. Fine. But, how does that help us deliver a rating system that gives us accurate ratings—the ones that the players really do have? Well, here is how the magic works. We use the true ratings of the players as a variable in the program. I call this the fixed rating. This implies that this rating does not change during the course of the simulation. I really shouldn't call it a variable, but when we create things in a program to use in computations, they are called variables. Constants are really what these true ratings are—at least as far as the program itself is concerned.

Next, we use these fixed ratings in a couple of ways. The first way is to decide which player is going to win the game. Because the program knows the fixed ratings, we can have it generate a random event such as a real number from 0 to 1 or throw a die with enough sides to get the appropriate resolution in the probability. Based on the difference in the ratings of the two players in question, we decide which player won the game. The following chart is what I used in my simulation.

Rating Difference	Cutoff on [0, 1]
< 100	0.5
100 to 200	0.4
200 to 300	0.3
300 to 400	0.2
400 to 500	0.1
500 to 750	0.0640
750 to 1000	0.0473
1000 to 1400	0.0255
1400 to 1800	0.0077
> 1800	0.0000

I used a hundred sided die to handle the first cases with a single decimal digit. I rolled my die and stored the result in a variable called j. Suppose we had players with ratings of 2430 and 2560. This is a difference in rating of 130 points. We see that this falls in the 100 to 200 category with a cutoff of 0.4. So, we roll the die and store the result into j. With the cutoff at 0.4, this translates into a 40 on a 100 point scale. So, we see if $j < 40$ (i.e. 0 to 39 inclusive). If it is, then the weaker player (the 2430) won the game. If it isn't, then the stronger player (the 2560) won the game. The rating system does not get to see the fixed ratings (the 2430 and the 2560). It only gets to see which player got the point. This process is how I conducted the simulation's millions of games.

I have a random number generator that I built using Steven Wolfram's Rule of 30 from his New Kind of Science book. I called upon this generator for each game in the simulation. So, I have every confidence that the required randomess is there for the rating system to work. I also used the random number generator to pair off the players for every round. I had to keep track of when a player was already paired off. Thus, there were 7500 games played during each round, and all 15000 of the players participated in exactly one of these games.

For the later cases in the chart above, I used a 10,000 sided die. Let's say we had a player rated 900 versus a player rated 2650. This is a difference in rating of 1750. That has a cutoff of 0.0077.

Translating this to the 10,000 sided die, yields a cutoff of 77. Thus, we roll the die and store it into j. Then, we see if $j < 77$ (i.e. 0 to 76). If it is, then the 900 rated player wins. If it isn't (which obviously is what happens most of the time), then $78 < j < 9999$, and the 2650 rated player wins.

Notice that for the simulated games I am describing here, we don't actually have a played game. Rather, we just simulate the random outcome of a 0 or 1 for each player to represent a played game. Thus, the details of the game are irrelevant as far as the rating system is concerned. We could be talking about chess, shogi, go, scrabble, backgammon, racket ball, abalone, ping pong, tennis, etc., etc., etc.

The second way we use the fixed ratings is to tell us after each round how well the rating system is doing. Since we know each of the true ratings, we can compute the value |true rating – calculated rating|. We can then see if this value is less than 100. I call a rating in spec if it is within 100 points of the true rating. So, I set up a counter to tell me how many players are still out of spec after each round. This leads us to ascertain the performance of this algorithm.

Elo's System

Chess uses a rating system for its players. This system is called the Elo rating system, named after Arpad Elo, who invented it. Elo's system has been in use for a number of decades now for chess and many other organizations where players can play one another. I have read that the Elo rating system is generally acknowledged to be accurate. So, in a real sense, the only challenge to it can come if a rating system can deliver the same ratings more quickly (in the sense of calendar days) or with less game data (in the sense of the quantity of the number of games played). But, we know that Elo's system has some problems. Yet, it is the established system, so uprooting it will require a demonstration of a far superior rating system.

Elo's system is an automatic system in that when a game is played, the results are fed into the formula genearating an automatic rating adjustment for each player. The two components of the PRA (The Basic System and The Boosting Sysem) are automatic in this sense, too. They are based on individually played games. No human intervention is required by the system. No checks and balances occur. Whether it is right or not is irrelevant. The Elo system and my Basic and Boosting systems are applied within their rules after every individual game played. The Tournament App (discussed later) is not implemented in this fashion. Rather, it is at the tournament level, and human intervention is required at certain times. The Tournament App can be set up to be automatic, but it has a component in it—called the Ratings Goodness of Fit Test—that should call for a ratings administrator to investigate should the tournament fail this test. The link to the Greatness of Chess Archive I supplied earlier has many many examples of tournaments passing and failing this test. Also, some tournaments are missing initial ratings, and the test is used on the ratings we do have to see if they can be used to get the rest of the ratings that are missing, but I am getting ahead of myself.

Algorithm Performance

The performance rating algorithm that I have designed is better than the Elo rating system used in chess. I have tested my rating system in a variety of ways through computer simulation. The results of this testing are what will be presented in this paper. Note though, that I have lost some of my data, and my simulation is now corrupted in some way when I moved into my current computer. So, when I get back to coding the simulation again (because I want a lot more data about it, and how other rating systems do against my population of 15000 players), I will fix whatever is corrupting it and build it into a beautiful data generating machine to assess how my rating system, Elo, and other rating systems perform.

My rating system is really two systems. I have a basic system that is very accurate, but it is slow in how many games is required to straighten out a population from a variety of disturbances from the ideal. So, to attempt to bring down the number of games required to accurately gauge the ratings of the players, I have designed a booster system. The booster system first measures the direction in rating of a player during a 20 game block of time. Then, it has some logic to tell it how much to boost the rating of the player in the appropriate direction.

To test how well this dual system works, I will perform a run on a standard population and another run on an elite population. The standard population is a bell shaped population of players with an average rating of 1500 and standard deviation of about 300. I have 15000 (fictional) players in my database with a fixed rating assigned to each one. The fixed rating is immutable in the program and is used only to inform the probabilities of each contest and to provide a count as to how many players have their rating more than 100 points away from their true rating.

I randomly assign the players to play one another in pairs, so 7500 games constitute a round. I have let the system run for 340 such rounds. This means 2,550,000 games are being tallied. A whole lot of random number generation and many calculations are being performed to simulate what should happen in the real world if we did this. We want to see how well the rating system performs under these conditions.

The initial ratings assigned to each of the players is called the seed ratings. If we seed the ratings exactly the way their fixed ratings are, then we have a stable population. We may want to do this to see how an individual player's rating that is incorrect performs against this stable population. How many games are required to get the incorrect rating corrected?

We may want to have the seed ratings systemically too high or too low from the fixed ratings to see how well the rating system fixes them. Again, the big question is how many games are required to get the population of players corrected? These disturbances from the ideal correct ratings are important to study because the initial rating we give to a player may then require quite a number of games to get corrected. If the rating system is fast enough, even a very bad initial rating won't require too many games to yield a correct rating for a player.

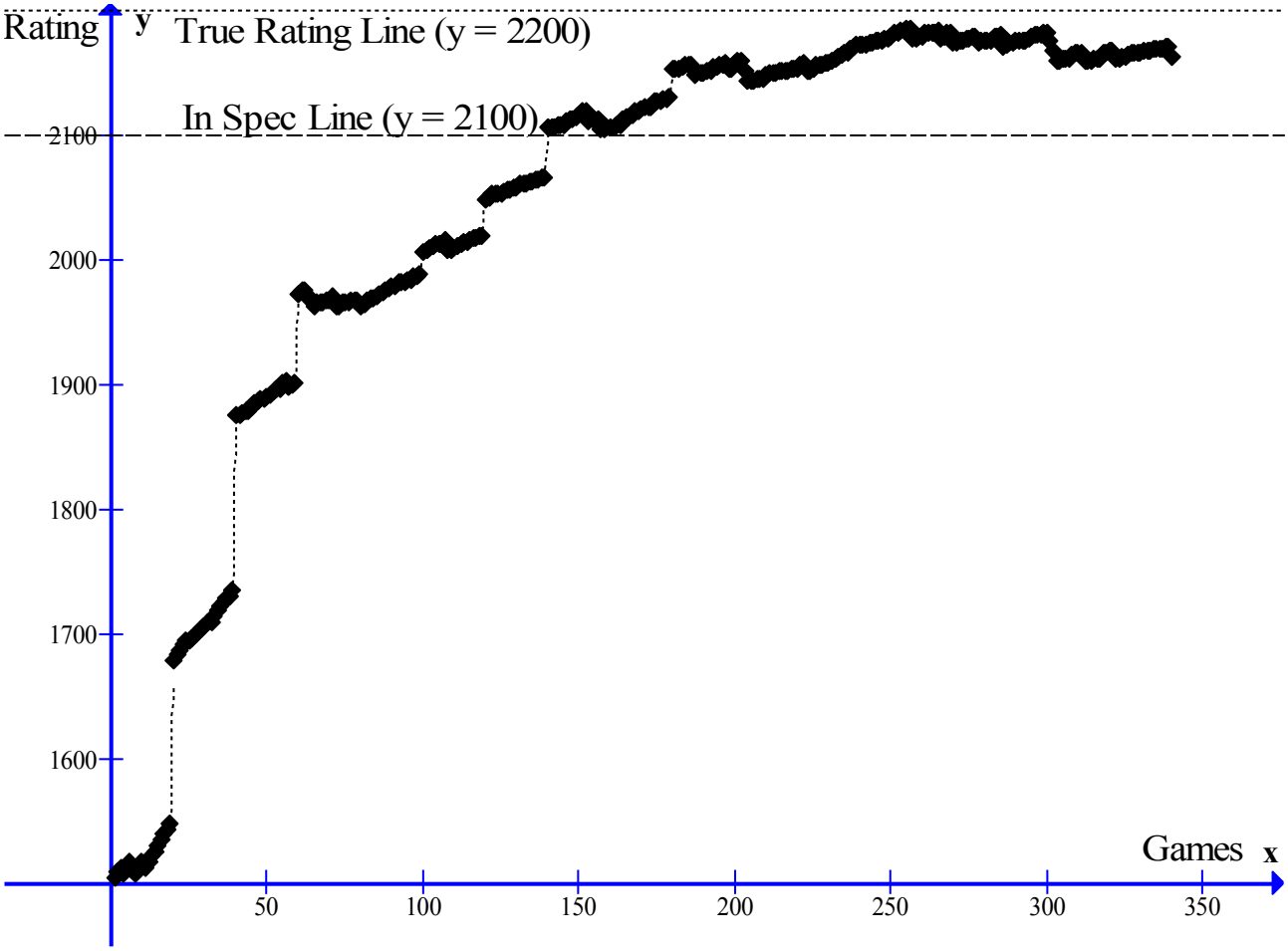
The worst disturbance possible is to have the entire population set to a seed rating of 1500.0. This is called the uniform distribution in statistics. With this disturbance, we want to see how many games it takes to straighten out the population. With bell shaped data, this is problematic because the players at the extreme ends of the distribution are many standard deviations away from the mean.

It could take quite a number of games to get them corrected. So, this is important to solve because we don't want to tell a potential future world champion that we will get his rating correct after he plays maybe 800 games! The rating system needs to do it faster than that.

Incidentally, Magnus Carlson, the highest rated chess player broke the 2800 rating barrier faster than anyone in history. This made news in the chess world. Readers of this document should take notice that my faster rating system brings down the number of games required by a player of Carlson's stature considerably. He may have been a 2800 player for a year or two before his rating finally reflected his true skill. This won't happen to a population of players using my rating system to rate them as we shall see.

As an additional side note, I have developed the Tournament App now, too. The Tournament App allows us to calibrate a player against an established field of players during a tournament with just

the data from that one tournament. In a game with tournaments that are bad for rating systems (like tennis with it's knock out style of tournaments), then just the basic and boosting systems should be used and not the Tournament App.

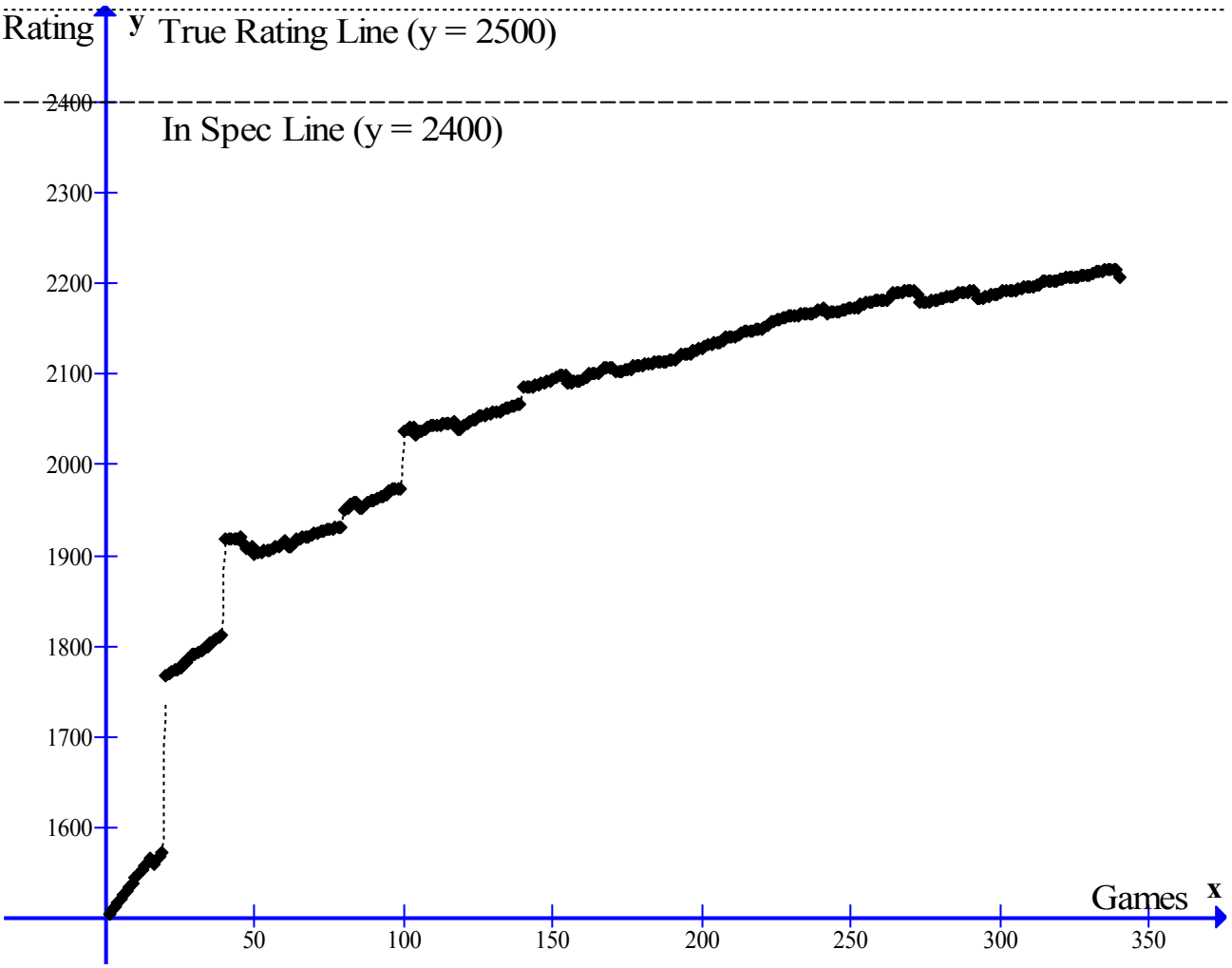


Ascent of a True 2200 Player on Standard Population

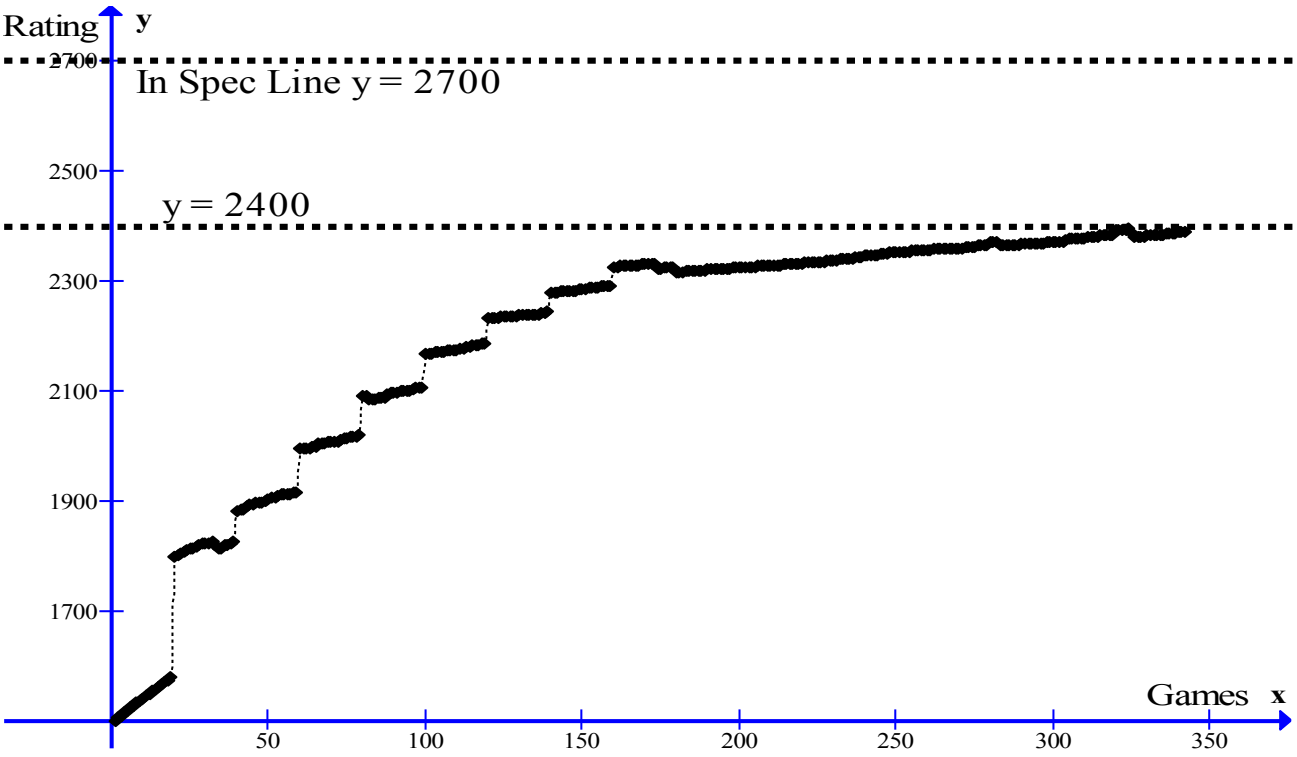
This graph is for the standard population from the initial state of chaos meaning all seed ratings were 1500 to start, including this player. The player depicted here had a true fixed rating of 2200. Observe the jumps in rating every 20th game. The boosting system is shut down after the 200th game for this particular run, but normally, they go in a reduced amount every 40th game after game 200 until game 320. So, the jumps are quite clear until then. Note that the jump at game 140 puts this player about 100 points away, so he is in spec at game 140. He inches upward after that, except for a downturn at about game 300. However, he stays in spec the whole way..

The graph on the next page is for a true 2500 against the same conditions. You can see that he is still well below the 2500 mark even after 340 games. It is hard for a 2500 player to find his true mark against opposition that is so far below him. The jumps help to boost him up, though. He would not rise even this high without them. The boost at game 100 puts our player above 2000. However, the quickness of rise is lost after that. We will see how a 2500 player rises better against an elite population shortly.

Many elite people have proclaimed the stupidity of the masses. Socrates said the bulk of the people are stupid, and he discussed the merits of the philosopher king. Emmanuel Lasker, in his *Manual of Chess*, asked if there was some force keeping the bulk of the people stupid. He laid out what he thought was a good system of teaching chess tactics and strategies with a time frame for each. Then, he considered this against the results that he actually saw out there in the real world. His



Ascent of a True 2500 Player on Standard Populations

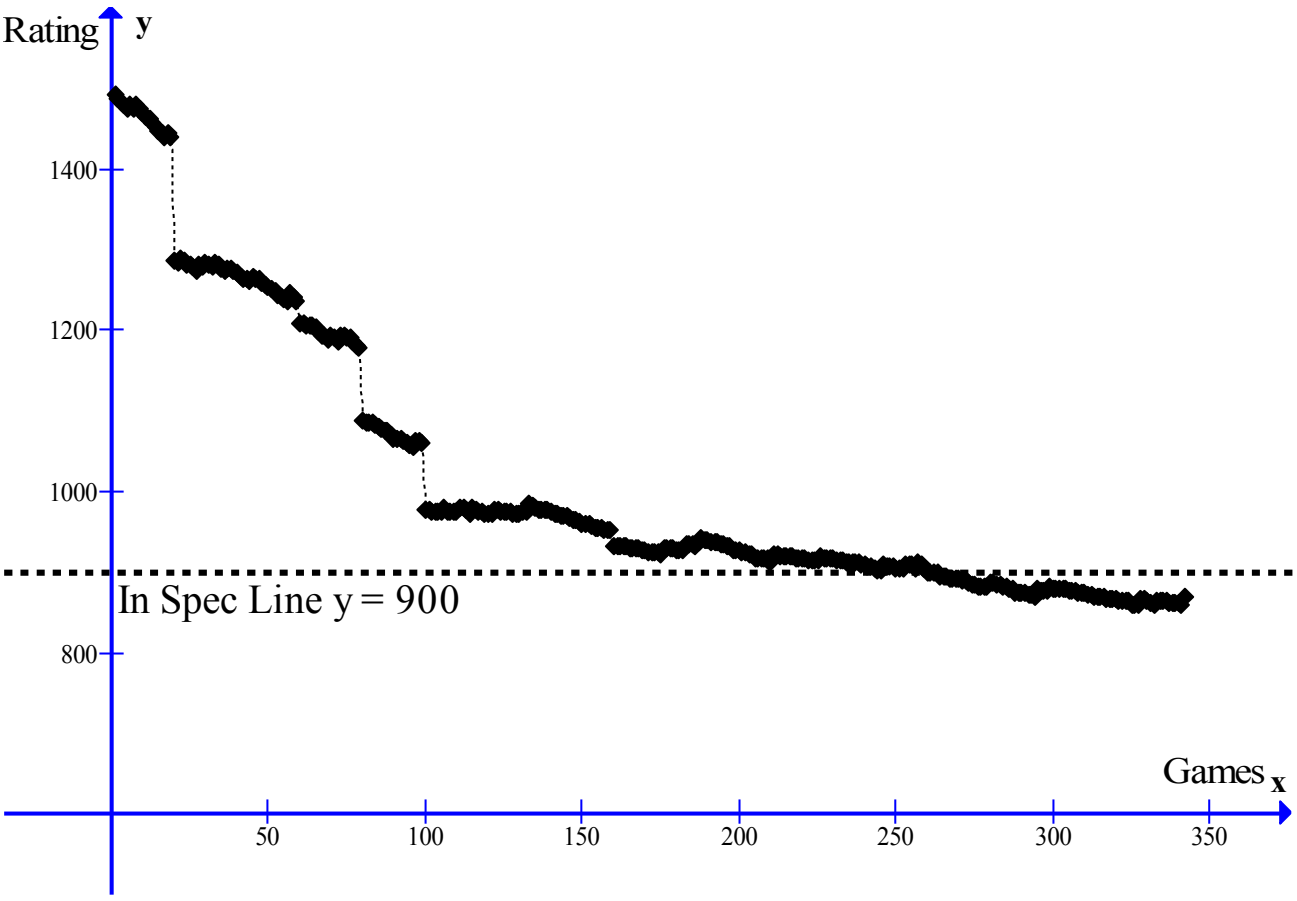


Ascent of a True 2800 Player on Standard Population

error is that he was elite, and his expectations cannot be obtained from the masses. Machiavelli also had dealings with the masses in which he wrote about unfavorably. Be that as it may, most of us are members of the masses in many respects. We may have an island or two of ability here and there, but it is rare for someone to have elite status in money, looks, intelligence, and charisma.

Here, in the above figure (on the previous page) is a 2800 player pitted against the masses from the initial state of chaos. Obviously, he rises better than the 2500 player did, but he is still well below 2800 after 340 games. The steepness of the ascent is quite low after game 170. We call this asymptotic behavior. He may eventually get to within 100 points of 2800 against the general population, but it will clearly take a great deal of games. This player is teasing the in spec line for a 2500 player near the end of the run.

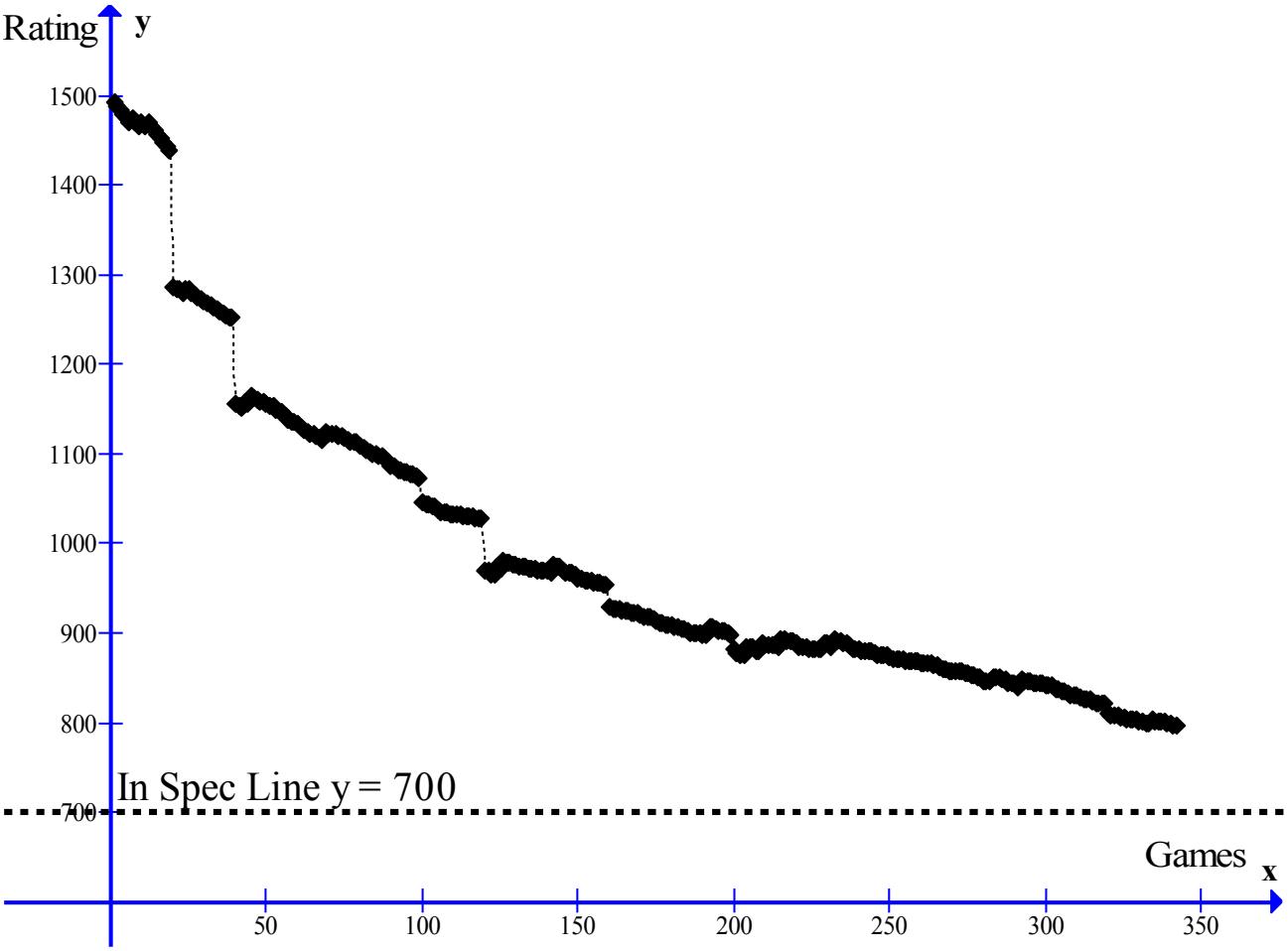
At the start of the run, all players had their seed ratings set at 1500.0, which is the average rating. This produced 11,039 players who were more than 100 points away from their true ratings. At the end of the run, there were 463 players out of 15000 who were more than 100 points away from their true rating. Thus, about 97% of the population was straightened out from the initial state of chaos. This is quite a remarkable feat, and the ELO system got to about 83% of the ratings fixed from a mild disturbance after 160 games. I don't know how well the ELO system will do from the initial state of chaos, but I will test it at some point and report on it.



Descent of a True 800 Player on Standard Population

The above diagram shows what happens to an 800 player. This player is a very bad player, and we can see from the initial seed rating of 1500 the descent. Three of the first four jumps downward are quite stark! It looks like at game 260 the player dips down below 900 for just a bit. Then, at game 275 or so goes below 900 again and stays in spec from there.

The next graph shows the descent of a true 600 player. Even after 340 games, this player is unable to reach the in spec line of 700. A 600 player is even worse than an 800 player, obviously. It shows that really bad players are as difficult to rate as the really good ones. The quotes above are not really true as we can see from this graph. The bulk of the people are not stupid. The bulk of the people are average. We can measure people of above average ability and below average ability against the masses. But, we cannot measure people who are extraordinarily good or bad against the masses.

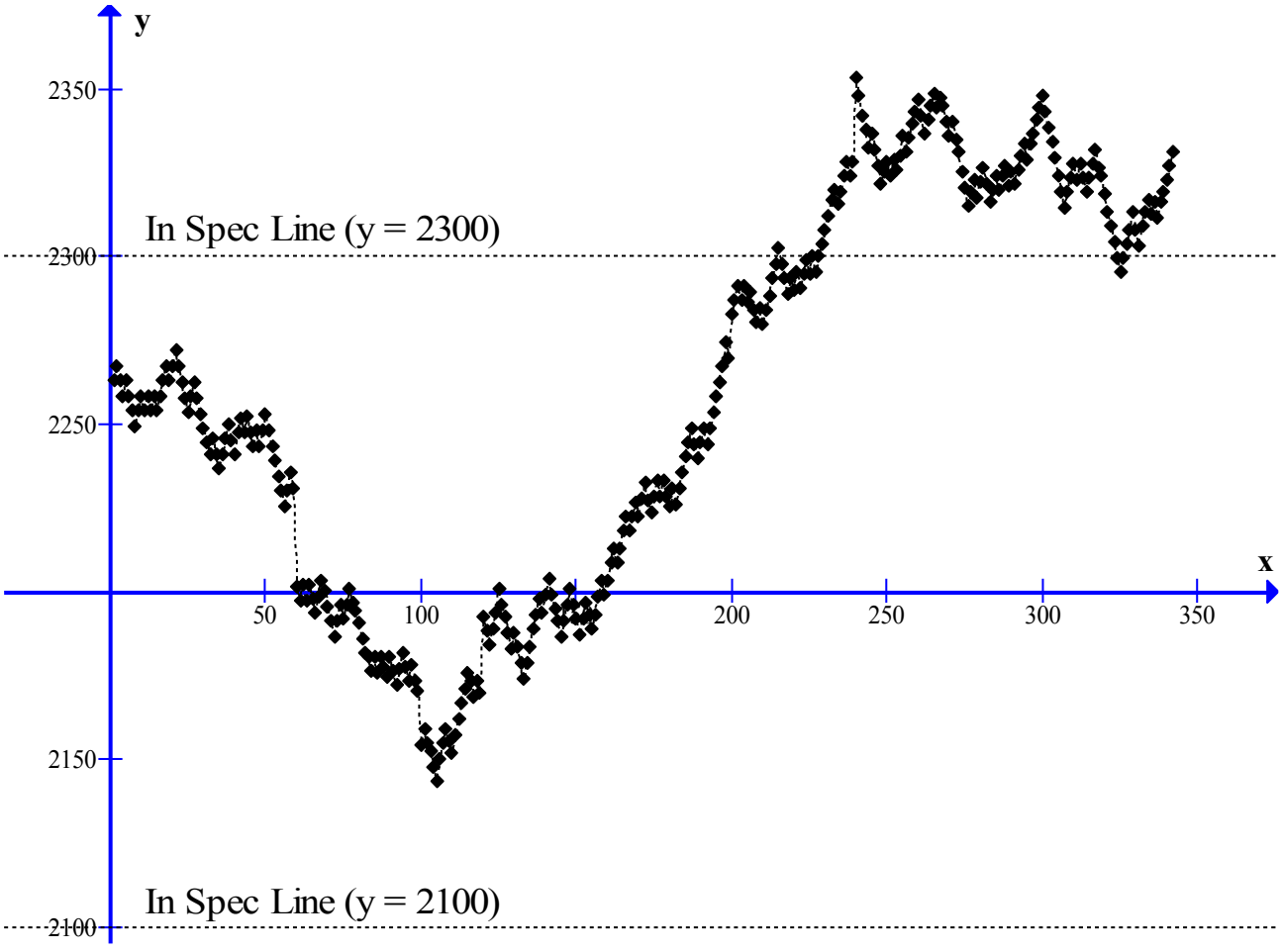


Descent of a True 600 Player on Standard Population

These graphs of the 2200, 2500, 2800, 800, and 600 players that I have shown here are not possible to see in real life. We can see graphs of player ratings over time, of course, but we don't know ahead of time what the player's rating is. With these graphs that I am depicting, we know the rating and can assess how well the rating system is doing to get its declaration of the player's strength correct. We can only surmise what the rating system is doing in real life. So, when we see a graph in real life, we must now ask, "Does this look like the real rating of a player?" If the graph is trending still, then it is not the player's rating. If it is sort of oscillating or fluctuating in some way, then it probably is hovering around the player's rating. But, the biggest problem about a trending graph is ascertaining where it will end. Look at the 600 graph. It is trending downward. We know it's a 600 player because the graph was introduced as such. If we didn't know that, can we say that this is a 600 player? This is why we wait for players to get whatever rating they are going to get from the rating system by playing more games against rated players. The USCF (United States Chess Federation) posts graphical images of the players in its database for anyone to see, but they don't try to predict the end-point of a trending player. Rather, they simply wait until the trend is over to see what the player's rating is.

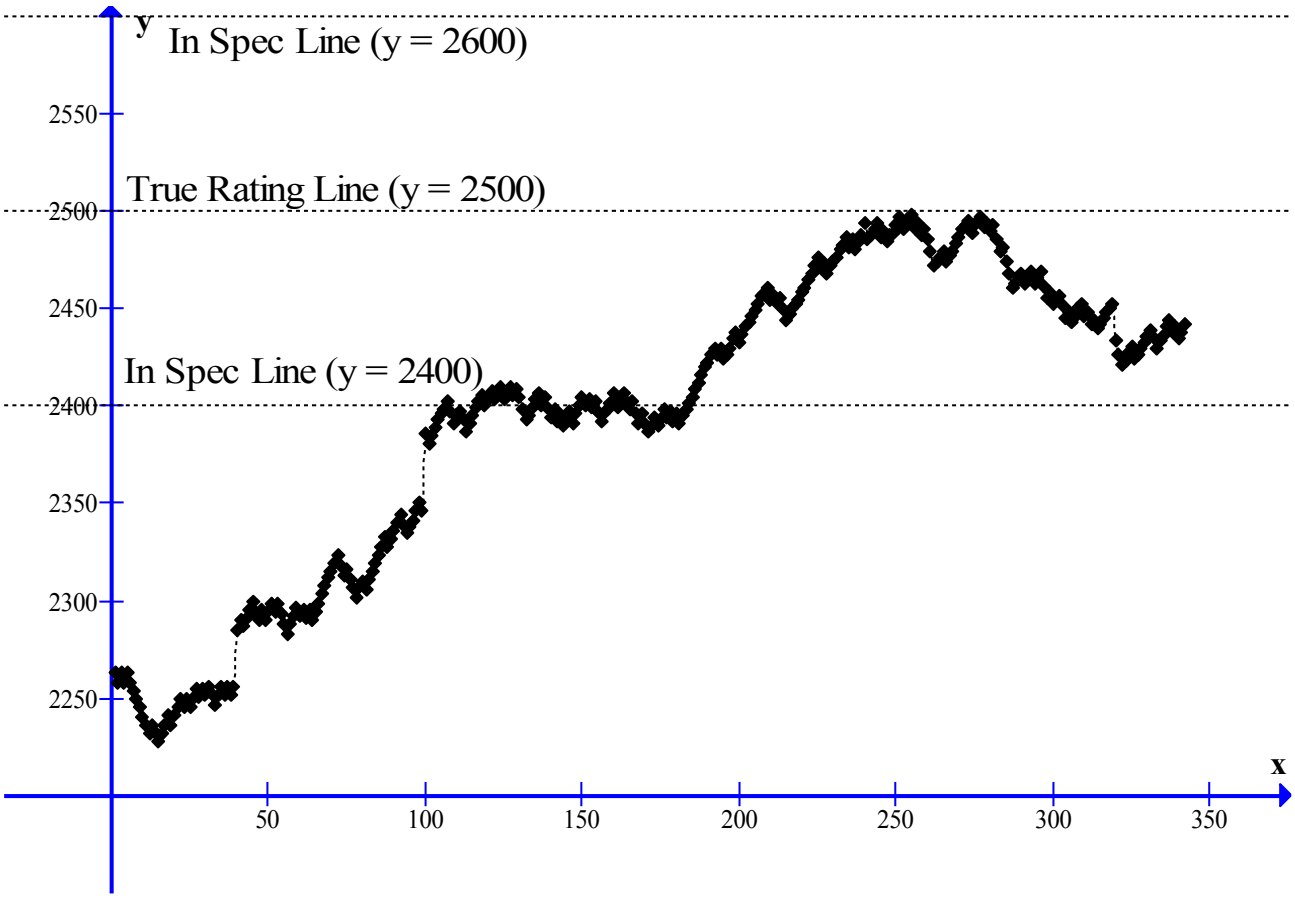
I have depicted graphs from the standard population of players. So, what will happen with our elite players against an elite population of players? I have a data set of elite players. It has 15000 members as well, and it starts at about 2150 and goes up to 2821. This population is the elite members of about a million regular players. So, this population is not bell-shaped at all. The shape starts high and just tapers down to near zero as ratings go along the values already stipulated. How will the performance rating algorithm do with this?

Well, here's the 2200 guy's performance over a 340 game span on the next page. The mean of the data set is 2258, so I started everyone at that rating. Thus, our 2200 guy is within 100 points of his true rating to start with. Here, the boosting system is jumping this guy around quite a bit. But, even at its worst, the guy is only getting to be about 160 points away from his true rating. This is about a 60-40 estimate in his chances against his true rating versus his rating system rating. The jerkiness is also due to many players at this same level are getting jerked, so it seems that it takes about 160 games to stabilize the population because there is no noticeable jump at game 180.



Rating Stability of a True 2200 Player on an Elite Population

Our 2500 guy can find his place among the elite on the graph on the next page. When he plays high level players, he can get an accurate rating. He actually starts out in the wrong direction, but he rights the course quickly. The boosts at games 20, 60, and 80 are negligible. Only the boosts at game 40 and 100 are good for him. He teases the lower in spec line at 2400 for a while until game 180 when he comfortably goes into his real rating region. The last boost at game 320 is in the wrong direction, but it is mild enough that he is still in spec. So, when elite players play amongst themselves, they can get accurate ratings with this rating system.

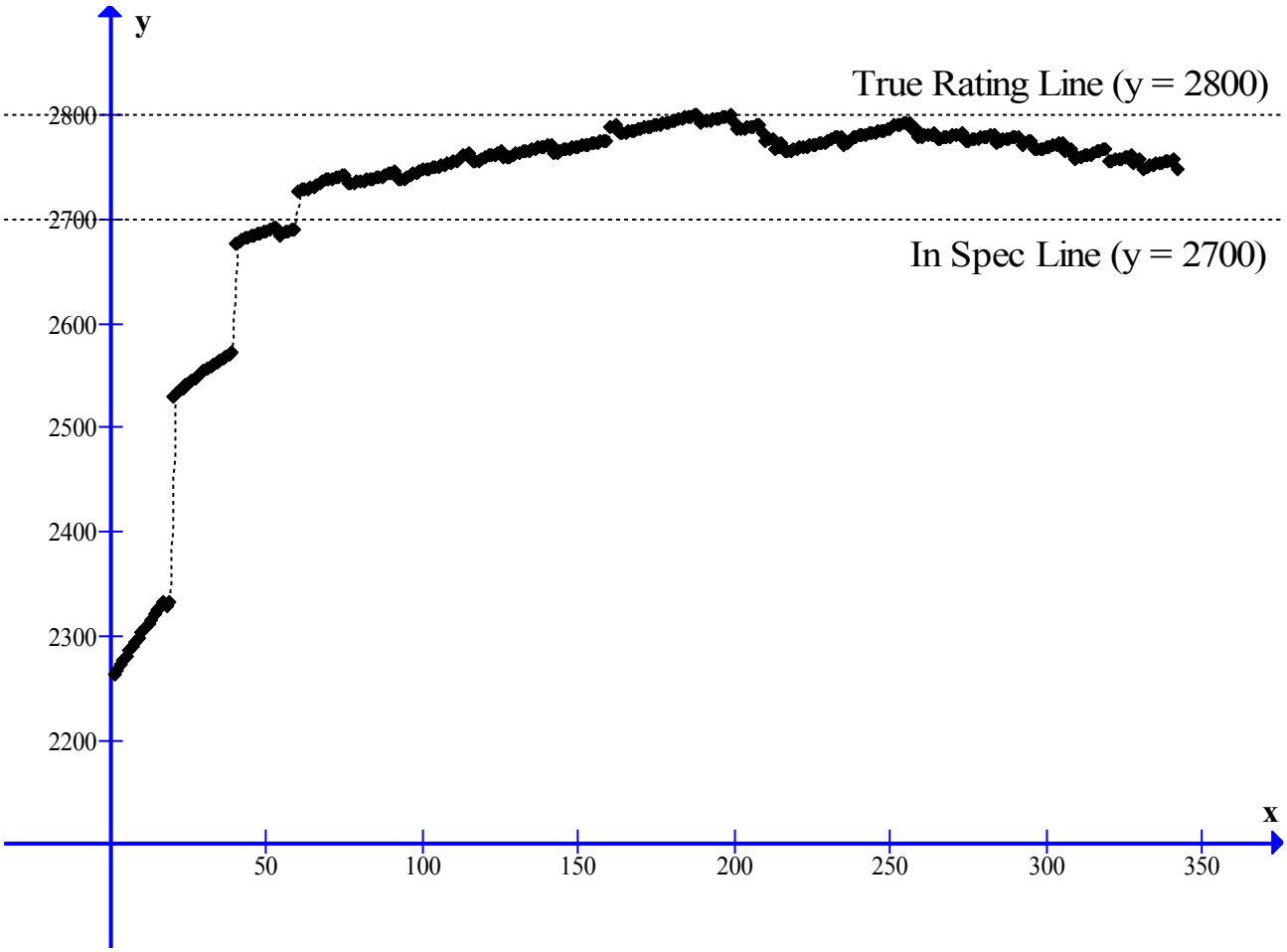


Mild Ascent of a True 2500 Player on an Elite Population

Our 2800 guy will be presented last on the next page. He gets two early jumps to put him over 2600 at game 40. The jump at game 60 puts him over 2700. Thus, with just 60 games, our guy becomes a member of the 2700 club. I had another run where the 2800 guy became a member of the 2700 club on his 268th game. So, there is some variation in how quickly the players get in spec. But, even with a long time to get to 2700 as the other example (not pictured), I am sure this is far better than what Elo can offer. I will get hard data on the Elo system against my population of 15,000 at some point when I fix my corrupted program. Also, I have a host of additional things I want to have it do. I will post on the progress of this as I go along.

The resolution of data presentation I am providing on my rating system (via the simulation) is going from the individual trees to the forest, so to speak. So, zooming out one notch from these individual player graphs, I will provide two charts. These charts provide one piece of information: How many games were played until the player became in spec (within 100 points of the true rating). Each row is a from a run of the simulation, and each column is headed by the rating of the player in question. I did this for the standard population as well as the elite population.

The following chart (also on the next page) is based on 44 runs of the simulation on the standard population. I set up some fixed ratings from 600 up to 2800 in increments of 100 to watch as the events unfolded. The ones with ratings below 900 and above 2100 had gaps in the data, which meant that more rounds than what I had the simulation set to were required sometimes. So, I just captured the inner group that is closest to the mean for this chart. What the numbers represent is a count of how many games were required for that player to get in spec. Recall that the players all started out at 1500. So, that column is filled with 1's as expected. The next ratings going out should be in spec as well if the player moves toward it right away. And, that is what we see with many 1's in these columns too. However, sometimes, large numbers are there. For example, the fourth run has



The Ascent of a Champion 2800 Player on the Elite Population

a 40 in the 1600 column. Also, observe that there is a 62 in the 1400 column. This is an extreme example of a rating that should be found quickly by the rating system needed an excessive quantity of games to render it in spec. Also observe the many ones place zeros in the table. This is due to the Boosting System operating at games 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 240, 280, and 320. Because many ratings get in spec at these rounds, we can see that the Boosting System is accurately boosting the ratings into spec. So, the Boosting System is well worth the trouble of implementing.

R900	R1000	R1100	R1200	R1300	R1400	R1500	R1600	R1700	R1800	R1900	R2000	R2100
76	105	80	33	80	1	1	3	20	40	120	60	140
280	182	60	60	20	1	1	1	20	20	60	194	169
211	160	212	40	100	2	1	13	48	20	40	134	100
128	160	40	40	40	2	1	40	20	94	60	40	103
313	119	51	100	20	1	1	1	20	20	95	100	121
105	121	68	80	28	1	1	1	20	20	40	60	100
80	96	44	20	40	1	1	1	20	36	40	100	140
80	100	80	48	40	1	1	5	20	81	40	60	140
80	46	140	80	100	10	1	19	20	80	73	140	257
80	100	60	20	20	18	1	1	83	140	240	45	312
120	82	40	88	100	2	1	1	40	40	82	120	100
280	182	82	83	40	2	1	5	20	72	40	97	97
272	40	80	20	40	6	1	7	40	49	120	60	233
80	187	40	53	20	1	1	1	20	32	88	140	80
128	280	83	82	60	1	1	1	20	208	40	60	172

100	60	125	40	40	1	1	1	60	40	40	120	100
182	260	40	56	64	1	1	1	63	100	60	60	118
264	182	62	40	20	12	1	1	20	61	100	140	120
100	140	152	20	20	2	1	11	40	20	80	120	80
189	120	40	20	20	2	1	5	20	44	100	60	228
120	213	125	40	98	1	1	1	20	60	60	80	200
200	160	60	160	40	1	1	1	20	91	100	170	340
238	100	60	142	20	2	1	1	117	21	160	60	160
100	76	60	40	29	2	1	1	40	20	80	51	92
160	194	40	20	74	1	1	17	20	40	53	175	160
100	60	80	52	64	6	1	3	60	20	81	100	285
209	122	40	52	140	1	1	1	40	40	80	246	203
80	186	100	74	20	2	1	1	99	40	40	60	144
93	60	64	20	25	2	1	1	20	120	128	134	165
100	100	60	110	20	1	1	47	89	20	137	120	160
100	297	60	140	20	1	1	1	57	20	60	60	123
140	317	100	20	28	62	1	11	40	39	80	223	80
80	100	20	20	20	36	1	1	20	40	65	160	289
119	60	80	23	20	2	1	1	40	60	120	100	155
103	88	120	40	81	19	1	3	20	40	81	218	164
142	67	100	128	40	8	1	19	68	40	40	120	337
249	344	140	20	40	1	1	1	52	40	43	100	480
120	107	40	40	20	1	1	1	200	283	131	111	140
210	146	100	100	60	4	1	26	20	40	40	60	80
257	119	61	97	20	1	1	1	130	40	40	80	140
60	60	40	84	20	1	1	1	107	201	66	120	60
185	160	94	140	20	1	1	1	52	40	60	252	85
91	120	40	140	60	1	1	3	100	20	100	119	100
134	60	52	20	139	1	1	1	20	20	40	81	80
6538	6038	3315	2745	2030	226	44	264	2085	2612	3443	4910	7132
148.59	137.23	75.34	62.39	46.14	5.14	1	6	47.39	59.36	78.25	111.59	162.09

Standard Population Table

I ran the simulation a number of times on an elite population as well to see how the rating system did on a population that didn't have the lower end involved. This would be more how it is in real life where the elite don't play against the average players. This population was clearly not bell-shaped and not symmetrical; however, the rating system worked just fine on it (Well, it appears fine until we look at the graph.).

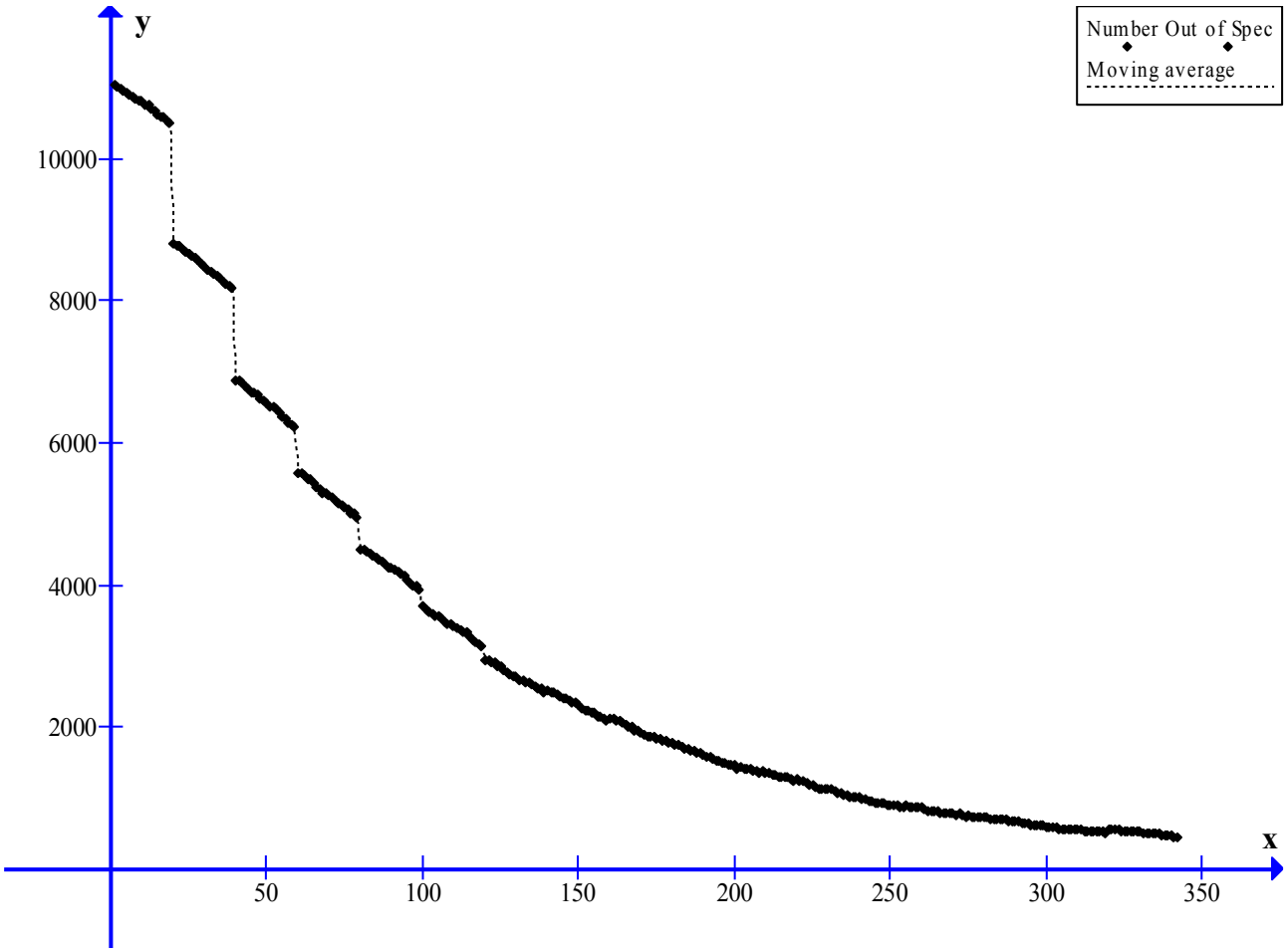
R1600	R1700	R1800	R1900	R2000	R2100	R2200	R2300	R2400	R2500	R2600	R2700	R2800
249	122	40	60	40	120	1	1	40	20	54	58	60
320	80	112	100	20	57	1	1	40	82	20	40	80
160	100	60	90	20	100	1	1	12	20	32	87	96
245	60	80	70	60	20	1	1	54	80	40	60	100
160	60	40	40	21	20	1	1	29	40	56	40	60
130	140	40	92	40	20	1	1	77	42	40	60	60
132	80	60	60	20	40	1	1	40	20	21	80	123
80	60	128	77	40	20	1	1	10	40	40	57	51
140	60	57	40	20	20	1	1	20	38	60	135	112
288	126	93	40	20	20	1	1	20	80	40	60	58
220	60	91	140	60	20	1	1	53	20	60	40	120

307	100	40	20	43	141	1	1	20	20	52	40	52
194	99	90	41	80	40	1	1	112	33	61	60	100
218	64	54	40	40	40	1	1	20	21	20	46	100
201	60	60	40	20	91	1	1	40	80	42	53	54
273	120	48	60	20	51	1	1	43	20	60	40	60
236	60	53	39	147	18	1	1	20	20	40	113	124
160	88	120	20	60	20	1	1	18	20	40	40	100
323	98	44	40	40	20	1	1	120	20	100	40	60
80	78	44	100	32	40	1	1	101	62	60	99	126
120	76	40	117	20	20	1	1	60	20	137	76	60
209	80	40	60	40	20	1	1	143	40	47	40	82
399	73	80	86	20	20	1	1	20	80	20	160	88
204	59	40	40	40	100	1	1	74	20	21	40	60
140	100	60	40	20	18	1	1	20	131	24	40	60
80	61	100	40	54	50	1	1	52	20	20	40	40
160	60	100	40	34	60	1	1	94	40	20	40	80
189	100	40	56	20	34	1	1	53	20	21	40	60
128	76	58	40	80	133	1	1	18	91	21	120	60
140	60	70	30	40	40	1	1	20	40	70	129	296
123	60	45	66	100	20	1	1	36	100	80	80	140
160	100	40	40	157	20	1	1	20	72	36	40	60
160	80	60	100	137	40	1	1	100	70	22	40	133
312	80	42	34	20	20	1	1	34	20	20	56	80
362	56	120	40	22	40	1	1	36	29	40	55	100
248	60	40	22	40	20	1	1	54	35	40	40	60
381	120	124	25	20	87	1	1	40	61	40	80	240
407	62	60	46	80	60	1	1	20	80	40	60	60
353	101	61	20	29	40	1	1	47	40	51	80	82
182	60	40	40	40	40	1	1	40	20	20	60	80
8573	3239	2614	2191	1856	1800	40	40	1870	1807	1728	2564	3617
214.33	80.98	65.35	54.78	46.4	45	1	1	46.75	45.18	43.2	64.1	90.43

Elite Population Table

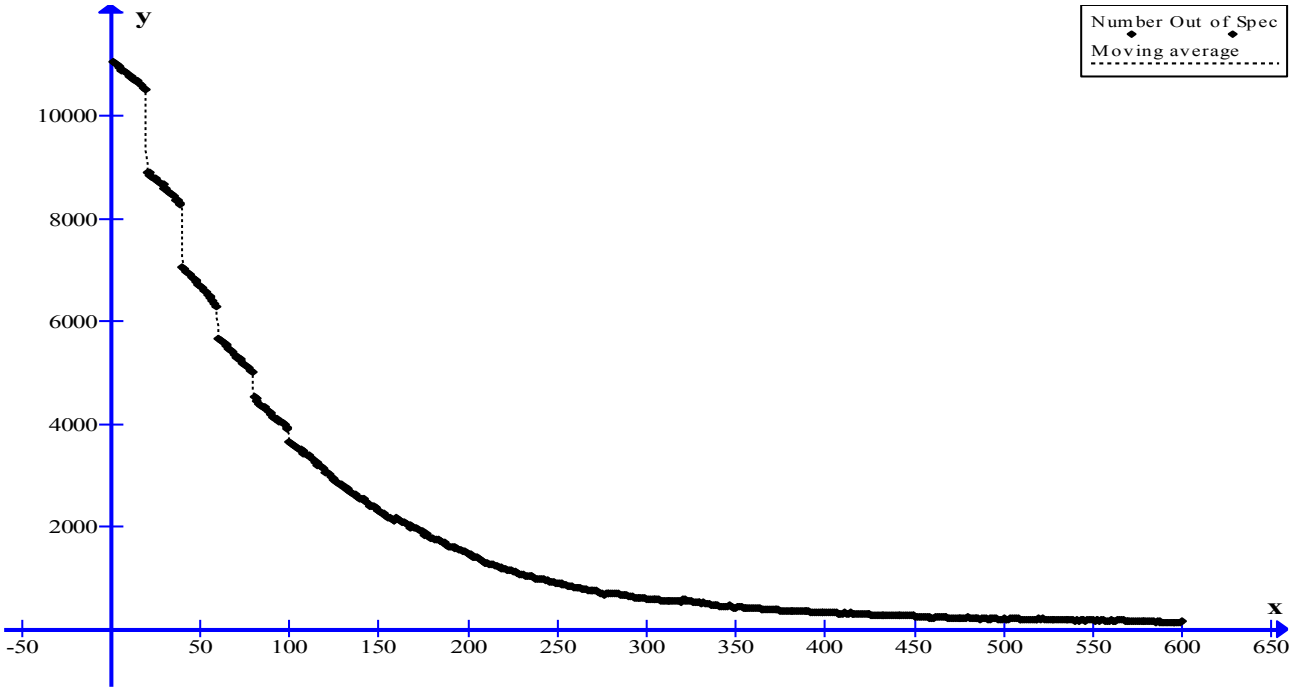
Moving out to the final zoom setting (the whole forest), we next will look at the algorithm's performance over time for the whole population. For the next graph, I ran the simution for 340 games, and plotted the number out of spec over time. Out of spec means at least100 points away from the true rating. From the initial state of chaos (all initial seed ratings set to 1500.0), we can see how rapidly the population is being fixed. This is a 340 game run. Observe that we are encountering an asymptotic decrease. Also note the gaps in the graph. The boosting feature is really an asset to the performance rating algorithm. It is leaping the population into being correct

If we increase our tolerance for being out of spec a little bit by declaring that being within 200 points is in spec, then at game 160, we get to about just 1% out because there are only 182 players out from another run of the simulattion. At game 400, within rounding, we get below 22 guys out of spec (it is 21), which means we are at about 99.9% of the population fixed by being within 200 points of the true rating. This is very impressive! This is with random pairings, remember. The elite will play many more games with the elite, so the rating system will do better with them. So, in real life, the system will perform even better than what is described here.



Number of Out of Specs for a Standard Population

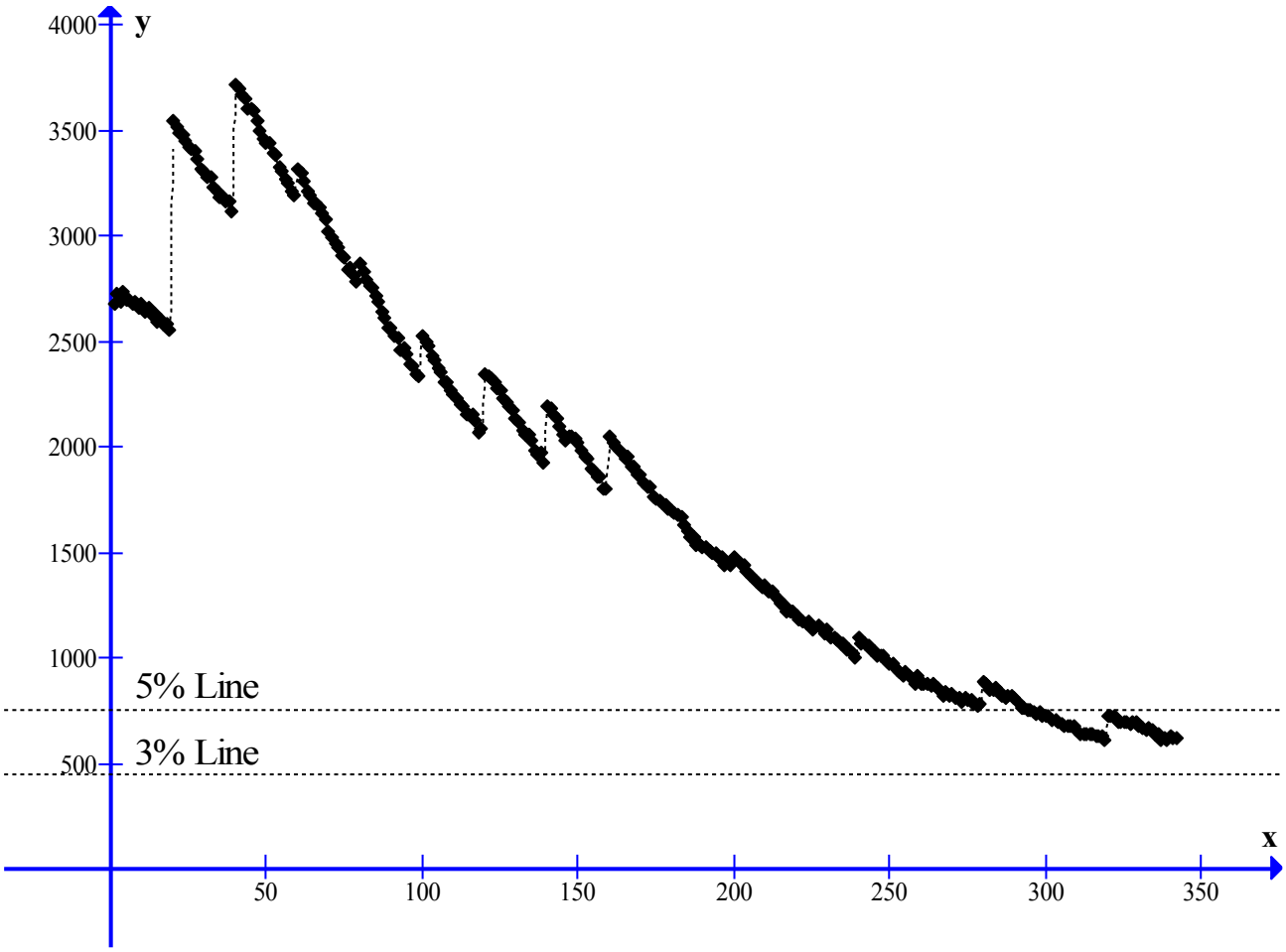
Next, this is another run of the simulation set to 600 games. We can see more of the asymptote, but by game 600, the population is hovering around 99% in spec.



Number of Out of Specs for a Standard Population

For the elite population, at the start of the run, only 2646 players were more than 100 points away. At the end, only 620 were more than 100 points away from their true ratings. This is about 96% accurate. So, it seems that about 96% or 97% accuracy can be achieved regardless of the type of group tested when the system is ran for 340 games. This success was achieved in spite of the Boosting System because it was clearly hurting things. However, this is the elite, and they play almost exclusively in tournaments. A protocol of the rating system is that once a player is calibrated in a tournament, then the boosting system for that player is turned off because the purpose of the boosting system (to get an elite player to an elite rating faster) will have been outgrown. Here is a graph of the situation for the elite population. The Boosting System is shown here to be a real detriment. But, we know that everyone in this population is elite, so the Boosting System is not needed because of that. In real life, we will not know that we have an elite population. It will probably be close to bell-shaped, but it may have some skew. So, in real life the Boosting System is definitely needed.

So, in conclusion, the performance rating algorithm is a success. High level players will not play the lower level players anyway. As their ratings climb, they will tend to play among their number. So, their true ratings will be found, and it will hover within the100 point range that defines the 50-50 odds. Players that descend in rating will also find their levels by playing against players at their levels. When this doesn't happen, a bad player can get a falsely high rating. For example, the lowest the 800 player got against the elite population was 1513.37.



The Number of Out of Specs for an Elite Population

So, the project is done. However, I will continue to work on this system to try to have a better solution against the asymptotic behavior. This is a real problem that Elo's system hasn't solved either as evidenced by Magnus Carlson's long climb to 2800, and his was the fastest on record.

[Note: This paragraph was written before I developed the Tournament App, so I decided to leave it here. Observe that I realized I got the Basic and Boosting systems done as well as they could be done, but I am not happy because of the asymptotic behavior problem was still clearly present.]

At some point, a website will be developed that hosts the wonderful game of Tines and Barbs. We have one in the beginning stages of development at www.tinesandbarbs.com if you want to inquire about its progress. Official ratings will be given that uses this system. This paper has shown how to implement the performance rating algorithm, so organizations that could use it may now benefit from it.

I have developed a new equation that has the ability to rate the great performances in chess's past and present. It can compare performances in spite of rating inflation. I call this equation the Greatness Equation. This equation uses the performance rating algorithm as part of what it computes. An entire chapter will be devoted to this equation. If you want to follow the development of this project, join my facebook group *Measuring Chess Greatness*. Go to the following web address:

http://www.facebook.com/groups/426121960802920/

I intend to maintain this page even after the Breakthrough Prize contest has concluded. This concludes the first major part of this paper, which is the performance rating algorithm, or the PRA. Next, we will discuss the Tournament App. We shall begin with the Ratings Law of Large Numbers after a brief discussion of how I prepare information to feed into the app..

Part 2: The Tournament App

It takes a bit of time for me to convert information on the internet about matches and tourneys into a form that I can use in my app. Here is how it needs to be so that I can use it. I will use Vishy Anand's Frankurt Chess Classics 1997 performance to explain how it works to get the information into a form needed to input into the app.

First, the program needs to know the name of the tournament, and I usually include the year. Note that at this stage, I have it be a single character string with no spaces. I use underscores to represent spaces.

Frankurt_Chess_Classics_1997

Next, the program wants to know how many players there are. So, in this tournament, there were four.

4

Then, the program wants to know the rating system used. In this case, it was Elo.

Elo

Next, the program needs a list of the ratings followed by the player's last name.

2765 Anand
2760 Karpov
2725 Topalov
2570 Lobron

After that, the program needs to know the rating of the number two player in the world at the time of this tournament or match.

2765

Finally, a list of the games (referencing player order above with the score of WHITE in the third position) is the last piece of information the program needs. As much as is possible, I try to have them in the order that they occured in real life. Sometimes, this is not possible—especially in older tournaments. But, when it is possible, I make sure that the order of the list be the same as it was in the tournament or match. I end the list with -1 -1 -1.0 to tell the app that this is the end of the data.

4 1 0.0
3 2 0.0
1 2 1.0
4 3 0.0
2 4 0.0
3 1 0.0
1 4 1.0
2 3 0.5
2 1 1.0
3 4 0.5
1 3 0.5

4 2 0.0
-1 -1 -1.0

This constitutes the input into the app. Then, I copy and paste this entire block of information into the app, and it processes the tourney. Then, I copy and paste the output file after the input part. From there, I copy and paste the input and output together and paste that into the *Greatness of Chess Archive*.

The Ratings Law of Large Numbers

The Ratings Law of Large Numbers is a special case of the Law of Large Numbers. In an elementary statistics class, one learns of two rules of probabilities—the classical one regarding the theoretical sample space formulation and the estimation one by repeated runs of an experiment. Both formulas are of the form:

$$P = s/n \tag{13}$$

In the sample space interpretation, P is the probability, s is the number of times that event occurs in the sample space, and n is the total number of items in the sample space. In the estimation formulation interpretation, P is the probability, s is the number of times the sought after event occured, and n is the number of trials the experiment is run.

For short runs, the estimated probability might not be too close to the theoretical probability. But, as the runs get longer and longer, the estimation gets better and better. This is what the Law of Large Numbers tells us. At the limit (as n approaches infinity), the two formulations will match exactly.

There are two corresponding formulations that are applicable in the Ratings Law of Large Numbers for matches. The theoretical formulation is

$$P = E(\text{Higher Rating} - \text{Lower Rating}). \tag{14}$$

Without loss of generality, we can place the higher rated player in the first position. The E is the expectation function for the Basic System. The estimation formulation is

$$P = \#\text{ points earned by the higher rated player / total number of games played.} \tag{15}$$

The probability for the lower rated player is, obviously, the complement (or 1 – P) of this in both formulations. In a match between two players, we get a short run estimate of the probability. From that, we can deduce the difference in their ratings. However, with the inherent problem of the consequences of short runs being inaccurate, we could be off in any computation regarding a difference in rating.

An example of this is when Steinitz, the world number one according to Chessmetrics, defeated Blackburne, the world number two, seven to nothing in 1876. The difference in rating, according to the estimation formulation would be

$$E(\text{difference}) = \text{Steinitz wins/total games} = 7/7 = 1, \tag{16}$$

which implies that the difference is $E^{-1}(1)$, which is 1800 points or higher. The -1 notation is to be interpreted as the inverse function of E. So, if Steinitz was rated 2800 as a world champ is by today's Elo ratings, Blackburne would be rated at 1000 points or less, which is a beginner. Blackburne was the number 2 player in the world at the time. According to Chessmetrics, Steinitz's

rating was 2730, and Blackburne's rating was 2648, which is a difference of just 82 points. Blackburne should have won slightly less than half of those games. So, Blackburne should have gotten 3.5 or 3 or 2.5, but we certainly would not expect for him to have gotten zero. Let us calculate what Steinitz should have gotten:

E(Difference in Rating) = E(82) = Steinitz's wins/total games
0.6 – sqrt((double)((diffT – 150.0)/-15000.0))

$$0.6-\sqrt{\frac{82-150}{-15000}}=0.532669967077586146977577427219 54$$

= Steinitz's wins / 7 ==> Steinitz's wins = 7(0.5327) = 3.73 wins (17)

This implies that Blackburne should have scored about 3.27 points, or rounding to the nearest half-game, 3.5 points. The nearest possible realization of this expectation is that Blackburne should have drawn the match. But, had they played 100 games, we would have seen Blackburne begin to level the match out with Steinitz getting maybe 55 to 60 and Blackburne getting 45 to to 40. More games than this would see even closer to expectation scores as the Ratings Law of Large Numbers exerted its influence on the match as it progressed..

Let us calculate the probability of this Steinitz sweep.

P(Steinitz sweep) = P(Steinitz winning first game and Steinitz winning second game and...and Steinitz winning the seventh game) = 0.5327 x 0.5327 x ... x 0.5327 (a total of seven times) = $(0.5327)^7$ = 0.01217 = 1.217%. (18)

So, Steinitz had just over a one percent chance of doing this. Because this is a small probability (a P-value, if you will), we would like to know if there was some kind of fraud that took place, or did we actually witness a rare event.

If you study chess history, you will find that Blackburne was a great player. And he did work himself up in the ranks to get to challenge Steinitz. There was a known hatred between them, and I have read that Blackburne (a large man) beat up Steinitz (a crippled man) once even. So, Blackburne would not have thrown the match. Therefore, it appears that we witnessed a rare event, which probably was enhanced by a psychological advantage of Steinitz over Blackburne. We refer to it as "getting in the head of the other player." Objectively, Blackburne was nearly Steinitz's equal, but with the psychological disadvantage of being swooned by the great man's chess exploits, Blackburne floundered and made small errors in his play that Steinitz exploited fully.

This sort of thing can happen in a short run. There are quite a number of lopsided matches in chess history to suggest the estimation rule is faulty for matches in general. Steinitz swept Blackburne. Lasker swept Bird. Tarrasch nearly swept Walbrodt. Capablanca swept Kostic. Fischer swept Tiamanov. Fischer swept Larsen. But, sometimes, it does work in matches quite well. It appears that the cuplrit is that one player may be intimidated by the other and makes moves that are not in line with their rating. But, there are many many matches that take place. I am only quoting the famous ones, so it may be in line with the total number of matches after all. When one only looks at the oddities, it can appear to skew the probabilities when nothing of that sort is going on. In the final analysis, then, we can say that from time to time in a short match between two fairly evenly matched players, we should expect a lopsided result or even an actual blow out.

In a tournament, though, I have found this intimidation factor to be substantially less. So many tournaments pass the Ratings Goodness of Fit Test, and this is true even when a player sweeps the field. Because the players in a tournament comprise a closed system, the effect is that the Ratings Law of Large Numbers exerts its power on all of the games of the tournament, which makes it a

longer run of games than for any one player alone. This is why in longer tournaments, the higher rated players are near the top of the pile, and the lower rated players are near the bottom. This tournament application of the Ratings Law of Large Numbers should surprise no one. If we see a violation of things like a rather low rated player doing extremely well in a long tournament, we might ask, "Are we seeing an extremely rare event if the ratings are correct, or are we seeing a quite normal event with the ratings simply being wrong?" I have an example of this in the Hastings 1895 tournament.

I have read accounts that everyone was surprised that Pillsbury won this tournament. I have also read about rapid increases in rating (called a quantum leap in rating), and Pillsbury must have had some such quick leap in chess ability for Chessmetrics to have him rated so much lower, but his real rating at the time of this tournament was about 2800, which is comparable to Lasker and the other elite players here. The adjustment in rating by boosting Pillsbury's and Pollock's ratings to their performance ratings is what I call a quantum correction in rating. The new ratings are Chessmetrics (PRA-corrected) ratings. Alternatively, we can call this a TRC Tournament Rating Correction or a TRA Tournament Rating Adjustment. I have this tournament processed in *The Greatness of Chess Archive* to accompany this paper in the submission to the Breakthrough Prize. This archive has a whole bunch of tournaments and matches processed with my app. It also includes two lists: the G-scores list, and the strength of tournaments list.

The tournament formulation for The Ratings Law of Large Numbers takes the form of a performance rating versus an initial rating. Thus, a performance rating is generated from the P-Zero scores, and the initial rating is generated by the rating system. We compare the two to see if they match. This is accomplished via the Ratings Goodness of Fit Test. We can have one or two players out of spec (but maybe not by too much) and the tournament will pass. Alternatively, we could possibly have everyone in spec but out by maybe 85 to 99 points each (i.e. everyone exhibiting pressure on their ratings), and this might fail a tournament. The test will only allow a certain amount of non-fitting before it trips the switch failing the tournament.

The Ratings Law of Large Numbers:

Let R_i , i = 1, 2, 3, …, n be the true but unknowable ratings of the players in a tournament. Then, $\widehat{R}_i$ is the initial ratings estimate for each of the players provided by the rating system. Given a random infinite set of pairings with the population of all players (or an equivalent, systematic set of pairings that cycle through all the players of the population), the $\widehat{R}_i = R_i$ for each value of i for the players in the tournament. Let r_i be the true but unknowable performance ratings of the players in the tournament. Then, $\hat{r}_i$ is the performance ratings estimate for each value of i for the players provided by the rating system based on the finite collection of game results fed into it. Given a random infinite set of pairings with the players in the tournament (or an equivalent, systematic set of pairings that cycle through all the players of the tournament), the $\hat{r}_i = r_i$ for each value of i for the players in the tournament. It is a given that $R_i = r_i$ for all i. The small variations in performance due to luck or tiredness on the part of a player or the opponents gets washed out in the infinite procession of games against one another, and at the limit when the games are infinite, the variation is zero. Thus, the performances are equal to the initial ratings at the limit. The ratings law of large numbers states that the $\hat{r}_i \to \hat{R}_i$ for each value of i for the players in the tournament. The fit gets better and better the more rounds the tournament lasts.

It takes quite a few coin flips to see the law of large numbers begin to exert its powerful influence. It is a 50-50 chance that any one coin flip will come up heads. Yet, over the course of many flips of a coin, the outcomes will balance out into about 50% heads and 50% tails. Likewise, it may take

quite a few games in a match between two evenly matched players to see a balance in the games. The many games played between Kasparov and Karpov has an edge to Kasparov as a case in point, but Karpov won enough of them to demonstrate he is not inferior by much.

But, the interactions between the players in the closed environment of a tournament lends itself well to the law. It takes comparatively few games by each player to bring the performances in line with the initial ratings (provided you use my rating system to detect this, not the official TPR's or tournament performance ratings). So, I have seen that as few as five or six games each is enough to do the job. The US Championship in 2013 was a Swiss style tournament with 24 players, and each player played nine games. This was a demonstrable success for my rating system with just nine games. However, I have quite a few tournaments in the archive that have players with less games than this, and my system produces the results just fine.

Currently, I am forced to use available ratings, such as Elo, Chessmetrics, or Edo as the $\hat{R}_i$'s . But, at some point, I will be able to implement my rating system on the huge databases of chess history, and I will then have better ratings information to process the tournaments with. I am interested in compiling a list of G-scores to see who was the best player in history. I am also interested in doing this for checkers, shogi, and go, among many many other game populations.

When I process a tournament, I develop the $\hat{r}_i$'s by using the P-Zero scores from the algorithm (The Performance Rating Algorithm's Basic System only, not the Boosting System). I regress the initial ratings on the P-Zero scores to obtain the performance ratings. Then, I use the Ratings Goodness of Fit Test to see how well the performance ratings fit with the initial ratings. Most of the time, the tournament passes. But, sometimes, it does not pass, and I have to figure out why the tournament failed.

If the tournament passes the test, then the ratings update in the customary way. But, if it failed, and I find no reason why it should have failed, then I am granted the authority to change initial ratings. Where does this authority come from? It comes from the fact that the P-Zero scores are the best measurements of the players during the tournament just concluded, and the Basic System is found to be the best measurer of ratings via the simulation I conducted. The technique of linear regression is beyond reproach to produce the performance ratings because of the Ratings Law of Large Numbers. The Ratings Goodness of Fit Test is unassailable as well because the modification I made (having a constant in the denominator) makes the test statistic chi-squared very robust in the face of changing initial ratings and a constant game data set. So, the only thing that changes the critical value is the number of players in the tournament. Thus, the only thing left to cause the failure is the initial ratings, which when changed to what the regression says it needs to be, the tournament passes.

The live example of the stunning success of this whole approach is the 2013 US Championship. I found three of the 24 players needed to have their initial ratings adjusted, and from there I produced a set of ratings for those 24 players. Over the course of the next 19 months, the test statistic obtained from the Ratings Goodness of Fit Test for my ratings in the E position (i.e. the "expected" versus "observed" in the chi-squared test) against the latest month's ratings for the 24 players in the O position produced a dramatic drop in the test statistic well into the exceptional fit region (the right tail) of the chi-squared distribution. The Elo system, thus, eventually basically got to the same set of ratings after many tournaments as my rating system produced from the data from just one tournament.

Thus, the method works, and it works because players perform at their rating levels during a tournament. The official TPR's can go sometimes up to 300 points away from the actual player ratings for the winners. Yes, it makes for great publicity, but it is a sham as a measurement. My

performance ratings (based on regression of the initial ratings on the P-Zero scores) allow the calibration of initial ratings, such as the three I did for the 2013 US Open. I identified Sevian and Troff as grandmasters at that tournament, but it took months and months before they both officially became grandmasters. Troff was first, and Sevian got his title about 18 months after the tournament. Kamsky's rating needed to be lowered. Who would suggest that Kamsky's rating be lowered? He won the tournament (after a play off with Ramirez), but my rating system said that he was out of spec by 115 points. Over the course of the 19 months of time, Kamsky's Elo rating has come down.

Being right about these ratings by picking the players who were out, and by how much they were out is so remote to get right by guessing, that it must be the work of a powerful rating system to be 19 months in front of the official Elo rating system. Not only does one have to guess which players will be GM's, but one also has to guess what the new ratings will be. Similarly, who needs to lose rating points and how many? My rating system nailed this with just that one tournament's worth of data. The Elo rating system required all of the extra data from the various tournaments that the players participated in during the 19 months that Elo took to hone in on the ratings established by my system with just that one tournament's worth of data. It is against the basic intuition to pick the two boys as grandmasters because they are just boys. It would be more conventional to pick a few of the players on the brink of 2500, the demarcation line, and say they were ready now for the next step. The two boys were much further out in rating to really believe that they were grandmasters. But, my rating system didn't know they were just boys. It didn't know the others may have been at the brink for a while now ready for that next step. It just did what it does and saw that Sevian and Troff were grandmasters based on their performances only, nothing else. The rest of the players on the brink by the way did not become grandmasters during the 19 months—only Sevian and Troff. The only player to drop any significant points was Kamsky.

I have some associated chapters in this document about the Ratings Goodness of Fit Test and the ways to adjust initial ratings if they are wrong. I have written case studies on tournaments as well that have missing initial ratings. Some tournaments have both incorrect initial ratings as well as missing initial ratings. Errors in the procedure of the Chessmetrics rating system produce these missing ratings, so I fill them in with my procedures as outlined in the case studies. But, the big point I am driving home here is that the Ratings Law of Large Numbers is the reason that we can use my performance ratings to calibrate initial ratings (from any rating system) with the incredible success seen with my Tournament App.

The Ratings Law of Large Numbers is a law I discovered while working on the simulation. I observed that the number of games won by the 2800 guy was about the same number more than that won by the 2700 guy, and that amount was also about the same amount as the difference between how much the 2700 guy won over the 2600 guy etc. There were slight variances from run to run, but it was consistent enough to cause me to notice it. I told various people I know about it, and I tried to use it in the simulation to get a better rating algorithm. Alas, it didn't help. It was simply too slow, so it remained a curiosity for a while. But, I did figure out where to place it in for its great contribution.

The rule of thumb we have is that a difference of rating of less than 100 points means that the two players are too evenly matched as to be able to call the winner ahead of time. Within 100 to 200 points, we get a 60-40 split in the number of wins for the higher rated player over the lower rated player. Next, we have 200 to 300 points difference yields a 70-30 split; and 300 to 400 points yields an 80-20 split. The largest rule of thumb difference is from 400 to 500 points difference yielding a 90-10 split.

This rule of thumb turns out to be the basis for the Ratings Law of Large Numbers. If I have a pair of players with a difference in rating of 150 points, that's right in the middle of the range for a 60-40 split. This means that the higher rated player carries that advantage over the rest of the players in the

field that they both have to play. Of the total points that they garner, 6 out of 10 will go to the higher rated player, and the remaining 4 out of 10 will go to the lower rated player. This does not count any games they play against each other, which merely has a probability of 0.6 for the higher rated player to win. Any one game in isolation is subject to probability, but a run of games is subject to expected counts based on the probability, which is what the Ratings Law of Large Numbers (and the Law of Large Numbers itself) reveal.

In small tournaments with few games, you can expect larger deviations from the law than in large tournaments. But, in larger tournaments, the law is working on all the players simultaneously; so, it becomes extremely unlikely that lower rated players will win such a tournament. I do have a lower rated player winning a large tournament—Hastings 1895, won by Pillsbury. Because of this law, his large out of spec from the Chessmetrics rating system, I corrected his rating with linear regression. The Chessmetrics ratings for the top five finishers at Hastings 1895 were the following:

2621 Pillsbury

2743 Chigorin

2860 Lasker

2823 Tarrasch

2763 Steinitz

Pillsbury won the tournament with a substantially lower rating than the other four. This kind of thing makes it very difficult to understand ratings in general. I will paste the regression and applicable part of the output for the Hastings 1895 tournament. From the original Chessmetrics ratings, we obtained the following:

Regression Analysis:

r = 0.77
Regression Eq: yhat = b0 + b1*x
Regression Eq: yhat = 2641.55 + 3.0227*x (19)
(xbar, ybar) = (0.00, 2641.55)

Residual Analysis (Magnitudes > 100 Become Suspect):

Rank) Name: P-Zero, Perf. Rating, Init. Rating, Residual
1) Pillsbury: 47.90, 2786.33, 2621, -165.33**LR**
2) Chigorin: 47.35, 2784.68, 2743, -41.68
3) Lasker: 43.29, 2772.38, 2860, 87.62
4) Tarrasch: 31.19, 2735.81, 2823, 87.19
5) Steinitz: 21.86, 2707.63, 2763, 55.37
16) Bird: -13.60, 2600.44, 2504, -96.44
19) Pollock: -22.20, 2574.43, 2462, -112.43**LR**

Thus, we have two large residuals—Pillsbury and Pollock. Observe that Bird is close to being a large residual, but he is in spec from the original ratings. We would say that Bird has upward pressure on his rating. I adjusted Pillsbury's and Pollock's ratings (both upward), which moved the regression line (the performance rating line) as a whole upward a little bit. I reran the app with these two new values and obtained the following:

Regression Analysis:

r = 0.88
Regression Eq: yhat = b0 + b1*x
Regression Eq: yhat = 2654.14 + 3.3705*x (20)
(xbar, ybar) = (0.00, 2654.14)

Residual Analysis (Magnitudes > 100 Become Suspect):

Rank) Name: P-Zero, Perf. Rating, Init. Rating, Residual
1) Pillsbury: 47.90, 2815.58, 2786, -29.58
2) Chigorin: 47.35, 2813.75, 2743, -70.75
3) Lasker: 43.29, 2800.03, 2860, 59.97
4) Tarrasch: 31.19, 2759.25, 2823, 63.75
5) Steinitz: 21.86, 2727.82, 2763, 35.18
16) Bird: -13.60, 2608.30, 2504, -104.30**LR**
19) Pollock: -22.20, 2579.30, 2574, -5.30

Observe that the residuals for all the listed players have improved, except for Bird, who is now out of spec. But, the protocols of the rating system are that we get only one shot a fixing the situation. So, Bird stays out of spec, but the value of r has gone up dramatically. We also pass the Ratings Goodness of Fit Test now. So, fixing Pillsbury's and Pollock's initial ratings enables us to see a much clearer picture of how this tournament went. I have just recently incorpated the concept of pressure into the system. I will have to see how it plays out if players exhibiting pressure get the benefit of a TRC or not.

Ultimately, the Ratings Law of Large Numbers is what is giving us the license to fix Pillsbury's and Pollock's initial ratings using linear regression. We will next discuss the P-Zero scores.

The P-Zero Scores

The P-Zero scores are actually easy to get. Have a list or array of float type variables for the players. Initialize them to 0.0. Then, as each game is played, let the Basic System adjust these scores as if they were initial ratings. As a tournament gets longer, the P-Zero scores gain more refinement over scores that are tied in points. This is because if a player beats another player with a higher P-Zero score in that point in the tournament, it is worth slightly more than beating a player with a slightly lower P-Zero score. These subtle differences are captured in the decimal expansion of the P-Zero scores. Look at any tournament in the archive for when there are tied players, and one will see small differences in the P-Zero scores for these players. I have seen only a couple of times when there are tied P-Zero scores as well. But then, we can look to the G-score to break the tie.

Thus, the P-Zero scores serve as the most accurate measurement of how the players did during the tournament. The P-Zero scores have some important properties. They are statistical measurements at the highest level—the ratio level. To be at this level requires a meaningful zero, and the P-Zero scores have this meaningful zero. A P-Zero score of zero means the player performed at the average level of the players in the tournament. A positive P-Zero score means the player performed above average, while a negative P-Zero score means the player performed below average. We can do stats on the P-Zero scores as well. The mean of the P-Zero scores is zero, but the standard deviation needs to be computed. I have a spot in the output where I report the z-scores, which is obtainable after we get the standard deviation. Because the P-Zero scores stand alone as the best measure of player performance during a tournament, they can wholly replace all of the tie-breaking schemes

ever invented, such as the Sonneborn-Berger score (which has a wikipedia article). The biggest complaint I have against the various tie-breaking systems is that they can result in a higher tie-breaking score for a player that scored fewer points. This is a major problem. I call this situation an inversion. My P-Zero scores have never suffered from an inversion through all of the tournaments I have processed in the archive, and quite a number of them did have an inversion when I had implemented the Sonneborn-Berger scores. At some point, I took out the code that computed the Sonneborn-Berger score because I had become disillusioned with it. I had reached the conclusion that the P-Zero scores were the best tie-breaking scores possible, so I am reporting that finding here.

Using the P-Zero scores, we can get the performance ratings, which are obtained by regressing the initial ratings on the P-Zero scores. However, the actual performance ratings are obtained by using the line

$$y = a + |b|x \tag{21}$$

instead of

$$y = a + bx. \tag{22}$$

We do this to keep the performance ratings in the right order should we happen to get a negative b. I have my app report that a negative slope was made positive when we actually do get a negative slope from the regression line.

We get the performance ratings by plugging the P-Zero score for a player into the regression line (or into the positively sloped version of it, if that is what happened). The performance ratings obtained in this fashion will be much closer to the initial ratings that what is called the TPR's or the Tournament Performance Ratings. The official TPR calculation is a bad calculation. It produces ratings that are far in excess or far below for the top performers and bottom performers, respectively. But, as good as the performance ratings are from regressing the initial ratings onto the P-Zero scores, we still need to test them for how well they fit. The Ratings Law of Large Numbers says that the initial ratings need to fit with the performance ratings, within a certain amount of tolerance, anyway. I have set my app to an alpha of 5% for where it marks the boundary of the right tail of the chi-squared test. But, I have had to make a small modification to the standard goodness of fit test in statistics. However, I have a whole chapter devoted to this test, which will be presented next.

The Ratings Goodness of Fit Test

Not so long ago (translation: I don't know exactly when this started), someone came up with the idea of the performance rating. It's very name suggests that a chess player's performance can vary considerably from tournament to tournament. I have read some articles about the greatest performance ratings ever. In one of the articles, Kasparov was the player mentioned with that honor for his 2001 Batumi performance of 11 out of 12 points for a calculated performance rating of 3000+ points, but I can't remember the exact number. Another article (wikipedia, I believe) stated that Karpov's 1994 Linares performance, and Magnus Carlson's Pearl Springs 2009 performance were the greatest ever (both over 3000 points as well).

However, can we not argue that Alekhine in San Remo 1930 is another contender? Nimzowitsch in Dresden 1926 (with Alekhine in the tournament) surpassed him, so how about that one as another contender? Capablanca in 1918 scored 10.5 out of 12, and his win against Marshall in that tournament is a very famous game (in chess circles, anyway). Capablanca in 1913 ran the table with a perfect score of twelve points in the New York Rice 1913 tournament. Recently, Topalov, in the Zug 2013 tournament, scored an impressive win by 1.5 points over his nearest competitor in a 12 man field where all the players were in the 2700 club. That has to count for something since some of the historical wins by previous great players included participants that were not of the first water, as is the case in all of the previously mentioned tournaments (including the two K's). As impressive as these individual performances are, they do not warrant the exaggerated ratings in which they get publicity in the chess press as "performance ratings" that in some cases exceed the players' historical ratings by several hundred points.

For the record, I have developed a better method of assessing great performances that I call the G-score, and G stands for greatness. I discuss it at length in another paper of mine. Using the G-score, I find that the greatest tournament performance ever (at least of the tournaments and matches I have processed) is Tarrasch's Vienna 1898 victory over a very strong field that included a 4 game playoff with Pillsbury.

All of the tournaments listed above are ones that I have processed. I have searched for the tournaments in which players over time have reportedly performed at their greatest levels. So, I am saying that Karpov and Kasparov do not have the greatest performances ever (at least of the tournaments I have processed). Over time, I may find a tournament in which one of the two K's outscores Tarrasch with the Greatness Formula. So, my proclamations here come "as of this writing."

Tarrasch suggested that maybe there should be two world champions: the tournament world champion, and the match world champion. Tarrasch had won a number of international tournaments at the time, so undoubtedly, he was referring to himself as the person deserving of that honor. Well, I sit here more than a hundred years later having processed the best tournaments by the greats that I have read about. None has surpassed the good doctor on the G-score. So, until further notice, I am declaring that Tarrasch has the greatest tournament performance ever, a proclamation that no doubt he would agree with.

In spite of my fine formula that computes G-scores, I continued in the attempt to find a performance rating calculation because I didn't really have one. I created some equations that I thought were correct, but it turned out that they failed. Eventually, I succeeded in finding this Holy Grail of equations, but the results surprised me. I found out that short term performances as in a single tournament are not different than a player's historical rating, which is earned over a long period of time by playing many games. Because of this finding, I am making the claim that players perform in individual tournaments at the same rating levels as their historical ratings. Thus, either I

am wrong, in which case you are free to stop reading me since I am incompetent. Or, the official performance ratings are exaggerated, and they are really in agreement with the player's official actual ratings if they were computed correctly. Note that I use the terms rating, historical rating, and actual rating to refer to a player's official rating. I might say raw performance or performance rating to refer to a player's performance in an individual tournament.

There is a great deal of publicity about performance ratings, so if I dare to be so **bold** as to be a voice of dissent against the establishment, I had better be bringing some heavy evidence to support my claims. Fear not, such evidence will be directly forthcoming (in this rather long paper). I have found a better way to compute ratings, and I have also found a way to find individual performance ratings for a tournament. I can run statistical tests of the ratings versus performance ratings of the players in a tournament, and I have found that they agree within reasonable statistical tolerance. So, the myth of the performance rating is busted, and I shall describe my methodology in the rest of this paper to show how I did it.

Paradoxically, I derived a surprising benefit. Performance ratings are never considered as official ratings because of their instability, yet with my performance rating calculation that I show in this paper, I run the Ratings Goodness of Fit test of the ratings vs. performance ratings to see if they fit. If the tournament fails this test, I can calibrate the out of spec historical ratings so that when the tournament is processed again with these fixed ratings, it passes. Thus, the unstable performance rating of old goes from being a hailer of greatness (by being way above even the greatest players ratings by well over a hundred points) to a calibrator of incorrect historical ratings. In other words, it is now the historical rating that becomes suspect, not the performance rating. Bear in mind, that is when you use my performance rating calculation procedure, not the official one (i.e. the TPR) that produces erroneous results.

For example, William Pollock is a better player than chessmetrics gives him credit for. I have two tournaments where Pollock's performance rating is significantly higher than the historical rating provided by chessmetrics. One of the tournaments (Hastings 1895) failed the test described in this paper, so I had the liberty of boosting his rating (called a quantum correction or a tournament rating correction). The other tournament passed (New York 1893), but Pollock's performance rating was nearly 100 points higher—consistent with the other performance rating. So, I am emphasizing that these new performance ratings are stable and accurate by way of this example. Below are the actual ratings for these two tournaments from chessmetrics, and my PR's (Performance Ratings) for William Pollock for these same two tournaments. Observe the consistency of these PR's. This shows that Pollock was a grandmaster. Also, in Belfast 1876, Pollock went undefeated by beating Blackburne and Burn enrout to winning the tournament. His G-score was an amazing G = 286.90, which is well over the 250 mark for being a threat to the world champion.

New York 1893: Chessmetrics had 2438, but my PR for him was 2533.51.
Hastings 1895: Chessmetrics had 2462, but my PR for him was 2585.23.

Next, I will say a few words about the PRA, which stands for Performance Rating Algorithm. This is the algorithm I discovered (not invented). This algorithm has two components: the Basic System, and the Boosting System. Together, as a rating system, this algorithm far surpasses the Elo rating system in terms of accuracy and speed. By speed, I mean the number of games that a player has to play to get an accurate rating. I have read that the Elo system requires about 30 games to become accurate. This is simply false. It takes more games than that as my simulations show, but my algorithm (the PRA) will whip a population into shape much faster than Elo. In the simulations, we start each new player at 1500, and as games are played, the basic system makes adjustments nearly like the Elo system does. But, the boosting system watches the basic system. It analyzes how the basic system is doing, and at regular intervals it makes dramatic adjustments if warranted.

Sometimes, it will not make adjustments if it sees that things are ok. I can take a population of 15000 players with normally distributed true ratings, and within 200 games by each player, this population is within 99% accuracy for 200 rating points. Elo does much worse than this because it can cause the population to drift over many games.

This system (with the two components of the Basic System and the Boosting System) is a self-contained rating system that works far better than the Elo system. But, it may drift out of control just as Elo's does. I have fixed that. Yes, you just read that correctly. I have fixed the rating inflation/deflation problem. It is such a silly fix that it required an obvious observation. We need to anchor the mean of the distribution at 1500. We could anchor it at any number, really, because the bell curve is capable of moving left and right with its shape intact. We could also compress it or expand it through simple transformation equations. However, Arpad Elo at some point in the past said that he thought the average player's rating should be 1500. So, I will not quarrel with that.

I will go further to declare that we can force the average rating to be exactly 1500 for the population of players in question. How can we do this when the equations will cause the drifting just mentioned against our best intentions? The answer is just to add up all of the ratings of the players in your population and divide by how many you have in the population. This value is your average rating. Subtract 1500 from this value. If you are left with a positive number, your average rating is too high. Subtract this residual from everyone's rating to bring the average back to 1500. If the residual is negative (i.e. your average rating is below 1500), then add this value to everyone's rating to boost the average rating back up to 1500. You can read about high level insight concerning where the drifting is occurring and all of that. Maybe some of that is going on, but it is difficult to see when we let the mean drift around. Anchoring the mean must be done first, and then you can start to look to see if more people are joining the 2700 club and other such things. Without an anchored mean, such analyses are meaningless.

Where does the drifting come from? There are several sources that I can see. One problem lies in provisional ratings. If a new player is given a provisional rating of 1200, then there is an automatic 300 point deficit in the total points in the system. Doing this a lot will bring down the average rating. This is why I start everyone at 1500. Another problem is people leaving the population of players. Players above or below 1500 in rating leaving will cause deficits or surpluses in the total points computation by their vacancies. For example, suppose you do have a population with the average rating at 1500, and a 2200 player leaves the population. This causes a 700 point loss in the point total, so now the average rating is below 1500. Exactly how much below 1500 is it? Well, that depends on how many players are now in the system. The final problem is that Elo staged a three tier K-factor into the equations: 30, 16, and 10. These cause disparities in the points added and subtracted from the winner and loser in an individual game of chess. In an isolated game, the disparity is small and negligible. However, when many games are going on across the whole system of players, these disparities can cause a shift in the bell curve.

These problems are why a whole system adjustment needs to be done fairly frequently, say, once per week. The problem is when the total points in the system is not equal to 1500*Q where Q is the number of players in the system. To those who deal with populations of ratings, you are welcome for me fixing the ratings inflation/deflation problem. But, I digress slightly. This paper is about The Rating Goodness of Fit Test. It isn't much of a digression, though, since we are trying to get accurate ratings with the goodness of fit test in the first place, and when the population drifts upward or downward, that doesn't help our cause. So, I needed to mention how to fix this problem. Next, let's talk about the elements of a goodness of fit test from statistics.

A goodness of fit test requires two things: something you are testing and something you are testing it against. You can have raw data or perhaps data that has been smoothed somehow as in a moving

average. This would be what you are testing, and it is the "observed" or O in the chi-squared test statistic equation. What you are testing it against is a known distribution or law or calculation or even another data set you have obtained like you did the first. This second thing is the "expected" or E in the chi-squared test statistic equation. Usually, it is an easy matter to distinguish the O from the E. In our situation, we have the ratings (which are data values obtained from many games played against a variety of opponents who also have ratings), and we have the performance ratings (which are data values obtained from just the few games played against just the players involved in the tournament). So, I have decided that the historical ratings (or just the ratings) are the O, and the performance ratings are the E. Now, I will show how I eventually got to the point of understanding that I had a goodness of fit test in the first place.

My PRA (Performance Rating Algorithm) measures the performance of each player during a tournament (as a P-Zero score). The players have a rating going into the tournament. I asked a logical question: Are these two related? I put my raw performance scores (the P-Zero scores) on the x-axis, and the ratings on the y-axis. Then, I did a regression line. Sometimes, *r* (the correlation coefficient) was high enough to say they were correlated (meaning it passes a hypothesis test for linear correlation). Sometimes, it was not. I believed that they should be correlated because the person who turned in a specific tournament performance was the same person who carried that historical rating. Thus, it seemed that it was matched pairs or dependent data of some kind. The issue at stake was did this person have a performance variation that could be widely different from the rating or not? The performance rating concept of chess implies that this individual performance did have considerable variation from the base rating. When a superstar "rises to the occasion" with a tournament performance rating of 3000+, it is a real crowd pleaser. We don't discuss the performance ratings of those players who were, say, 200 points below their ratings. But, these overblown ratings and terribly low performance ratings really muck up the situation. When we have these extreme fluctuations like this, the associated *r* will fail the statistical test. This happens when there is a wide enough dispersion in the historical ratings. Wide dispersion will pass the linear correlation test. Narrow dispersion will fail it. Why were the tournaments doing this? It was so frustrating.

Well, I hit upon the idea that the predicted $\hat{y}$ of the regression line was the performance rating concept I had been looking for, but because I had some tournaments in my database that had *r*'s that were low, it looked as if the high variation hypothesis for individual performances was right. But, I didn't give up!

Later, I had an epiphany. The $\hat{y}$ of the regression line (the predicted value) actually was the expected value (the *E*) of the player for a goodness of fit test. Thus, $\hat{y} = E$. I then formulated a hypothesis test that the players all performed to expectation using a goodness of fit test. The chi-squared test statistic was

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{23}$$

for all the players in the tournament. I used alpha at 0.05. The observed was their ratings. The expected scores were the predictions from the regression line based on the individual tournament performances (the P-Zero scores). So, I have a critical value that can be looked up, and a test statistic. It looks like we have a goodness of fit test now, but not quite.

This almost solves the issue of the performance rating vs. the rating. The last remaining problem involves the ability of the test statistic to move around when we add or subtract a certain amount from the O's. Given a set of ratings, we can have these players play one another in a tournament.

This generates a set of P-Zero scores. Now, suppose we take this set of players and add a set amount to their ratings, say, K. Then, the critical value stays the same because we have the same number of players. However, the test statistic moves. The numerator stays the same because the O – E will become (O + K) – (E + K), which is O + K – E – K. So, the K's cancel out. But, the denominator of E becomes E + K. So, this shift of K units to everyone's rating causes a change to occur in the test statistic. It would be great if we could avoid this.

So, I figured out that I needed a constant K in the denominator, but what should it be? (Note: the K mentioned earlier is a shift of K units in the original ratings. This new K is a constant I am attempting to find to be sitting in the denominator alone without adding an E. They happen to be the same letter. Sorry about that. K is a common choice for a constant in math.) Before too long, I realized that I should set my app to give me a confidence interval for each tournament I processed. I could look at these confidence intervals to ascertain what the best value for K needed to be. Here is a list of the tournaments I processed with these confidence intervals. This list has been placed in chronological order. Note: Some of these confidence intervals may have changed a little bit because I modified my procedure slightly. But, these are the original confidence intervals I looked at to decide what my K needed to be. Because of this, I am leaving them here as they originally were.

Berlin 1881: K $\in$ (1717.7, 5673.4)
London 1883: K $\in$ (796, 3022)
New York 1893: K $\in$ (1921, 7291)
Hastings 1895 (Chessmetrics before fixing): K $\in$ (2965.3, 8357.9)
Hastings 1895 [fixed]: K $\in$ (1577.3, 4445.8)
St. Petersburg 1895: K $\in$ (742, 16476)
Vienna 1898: K $\in$ (1519, 4525)
NewYorkRice 1913: K $\in$ (1134.5, 4305.5)
New York 1918: K $\in$ (1570, 12096)
Hastings 1921-2 (Before fixing): K $\in$ (3481.3, 22595.3)
Hastings 1921-2 (After fixing): K $\in$ (2232.6, 14490.7)
Dresden 1926: K $\in$ (1765.8, 8984.8)
San Remo 1930: K $\in$ (1541.2, 5305.4)
*AVRO 1938: K $\in$ (137.7, 894.0)
U.S. Open 1963-4: K $\in$ (831.2, 3574.7)
USSR_Championship_1964: K $\in$ (2237.7, 6667.2)
Interzonal 1970: K $\in$ (1562.6, 4198.5)
Tilburg 1989 (Fide ratings): K $\in$ (427.5, 2774.9)
Tilburg 1989 (Chessmetrics ratings): K $\in$ (575.0, 3732.1)
Linares 1994 (Fide Ratings): K $\in$ (1039.2, 3944.3)
Linares 1994 (Chessmetrics Ratings): K $\in$ (1647.1, 6251.4)
Las Palmas 1996: K $\in$ (187.6, 1812.8)
Frankfurt 1997: K $\in$ (761.7, 16918.4)
Tata Chess 1999: K $\in$ (1105.3, 4195.1)
Linares 2001: K $\in$ (420.6, 4065.3)
*Tal Memorial 2009: K $\in$ (183.2, 931.9)
Pearl Springs 2009: K $\in$ (277.9, 2685.8)
London Classic 2012: K $\in$ (924.4, 5246.1)
Tata Chess 2013: K $\in$ (1212.6, 4602.3)
Candidates 2013: K $\in$ (816.8, 5301.6)
U.S. Open 2013 (Original): K $\in$ (3587.9, 9640.3)
U.S. Open 2013 (Repaired): K $\in$ (2112.2, 5675.3)
Alekhine Memorial 2013: K $\in$ (666.9, 3393.3)
*Zug 2013: K $\in$ (340.5, 1464.6)

Norway Supertournament: K ∈ (1014.4, 5161.6)
Sberbank Chess Tournament: K ∈ (742.1, 3775.9)
Thesseloniki Greece 2013: K ∈ (409.9, 1762.9)

When I used the regular E (the $\hat{y}$ prediction from the regression equation), nearly all of the tournaments passed this chi-squared test for goodness of fit. This proved that the chess players of the past (and the current ones participating in recent tournaments, such as the Candidates Tournament) performed at their levels. But, I needed to pick a K from all of these confidence intervals that gave me similar results. The ones with an * are safe in that their CI's are so low that any reasonable K will get them to pass. So, I looked at them collectively and found that a K in the 2200 to 2600 range will get the same tournaments to pass and the same tournaments to fail. I chose 2500, since it is a perfect square and is the most easy to remember of these. The value of 2500 also happens to be the cutoff rating for a grandmaster. Thus, I now have my unvarying test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{2500} \,. \tag{24}$$

Having an unvarying test statistic gives the test an incredible robustness that is very important. The value of this cannot be overstated, and this test is now the litmus test for me when I process a tournament. The first thing I look at is whether or not the tournament passed this test.

So, now I have my goodness of fit test. Because I modified it, I get to name it. I call it the *Ratings Goodness of Fit Test* since it is used to check to see if the ratings fit with the performance ratings. This is a contribution of mine to mathematics. To those who appreciate this, you are welcome. I had a friend of mine get a couple of points named after him in a branch of mathematics called Triangle Geometry. They are the Garcia point, and the Garcia-Feuerbach point. I posted this to a reach of about 20,000 people on facebook in the groups I am associated with. After about three days, I went back to see what traffic I got on those posts. It was disappointing. I got a one sentence comment about the online encyclopedia link I provided to the triangle centers. I also got seven likes. These 20,000 people were ones who had liked or joined these math pages on facebook. So, I know how unresponsive the world (the math world) is to new discoveries in math. However, some of the readers of this document will appreciate what I have done here, and it is those people I acknowledge and appreciate. In the future when tournaments are processed with this test, and we detect a cheater because of it, they will just say, "Well, that math stuff I guess is good for something!" Oh well, my rant is over. Well, I will say further that my friend Emmanuel Antonio José García is finding more points. Thus, he is active in the branch of mathematics called Triangle Geometry.

I have now found that players did not rise to the occasion. They did not choke. Some guy posted a list a "chokes" on the internet. I put in the words "Korchnoi chokes" to get it. I don't know the methodology being used, but even if valid, it didn't have anyone with more than two games out. Thus, a game and a half was the worst "out of expectation" on this list. This is my findings, too, although the guy posting this list is just showing his out of specs and calling them chokes. I am taking the whole tournament into account with this chi-squared test. It allows for a little variation (small chokes, if you will) to still be able to pass the test. So, on the whole the players played to their abilities. It was both a thrill to have discovered this and a responsibility to share it. So, the crowd pleasing 3000+ performance ratings are false. Sorry to those who enjoyed those grand numbers. The big flops where a 2700 club player dips into the 2500's for a performance rating is also false. So, if you get nothing else out of this paper, just know that the elite players are fighting to their abilities, no more and no less.

After doing more tournaments, I found that Hastings 1895 had two players who performed significantly better than their ratings (as found on the chessmetrics website). Hastings 1895 failed

the chi-squared test. I wound up implementing my refined system on this tournament. Pillsbury and Pollock both got their ratings deleted from two separate Ratings Goodness of Fit Tests (see the archive). Along with three missing initial ratings, Pillsbury and Pollock had their initial ratings found with the following table using the regression line that finally passed.

Name	P-Zero	I.R.
Pillsbury:	47.90	2818
Burn:	-8.51	2631
Pollock:	-22.20	2585
Tinsley:	-26.22	2572
Vergani:	-64.92	2443

Then, the tournament passed. I feel justified in doing this, and I will do it for players in my Tines and Barbs (the board game I invented) tournaments, too. Players perform at their levels. I have seen it work too many times, so the few times where it doesn't manifest must be so because of the initial ratings being incorrect. This is why I fix the historical ratings. I have corrected Elo and chessmetrics ratings by this method. Will a PRA rating get corrected? Well, that remains to be seen. The P-Zero scores (the E's) are scores obtained from the PRA for just that tournament, and the ratings of a population of players from the PRA would one day be historical ratings (the O). Those could be subjected to this test. Peering into the future, I would say that quantum corrections probably would occur, but they should be more rare than with Elo or chessmetrics populations.

By extension, many individual sports endeavors such as tennis, I am confident, share this. One merely has to get the correct rating system and individual tournament performance measurement system set up to see it. Then, do a regression analysis and a chi-squared test to get the results. So, we shall now get into the details followed by some examples.

The regression equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \,. \tag{25}$$

So, renaming these, we can get the performance rating equation

$$\text{PR} = \bar{y} + |\text{m}|*\text{P-Zero}. \tag{26}$$

In words, the performance rating is equal to the mean of the ratings in the tournament plus the absolute value of the slope of the regression line times the P-Zero score. What is fixed for an individual tournament is its mean ($\bar{y}$) and the slope of its regression line (m). Thus, the P-Zero scores alone are what determine the differences in the player's performance ratings.

Now, I am guilty of a bit of sleight of hand. The slope of a regression line can be positive or negative. I have forced mine to be positive only with the absolute value bars around the m when I constructed my performance rating equation. Why did I do that? It's because it is possible for this slope to be negative, but only when the initial ratings are close together. This happened in a recent tournament (discussed a bit later), and the PR's were going down as the P-Zero scores were increasing. Statistics tells us the best prediction of $\hat{y}$ is $\bar{y}$. However, that won't work here since the PR's are supposed to decrease for players who performed worse, not increase or stay the same. So, I decided to retain the magnitude of the slope from the regression line. In the instances where we do get a negative slope, it's magnitude will be small. So, this is really only a slight deviation from the standard theory. And, one very important point is shared by a negatively sloped PR line that gets

changed to the positively sloped one: the centroid. Statistics theory states that the centroid $(\bar{x}, \bar{y})$ is a point on the regression line. How is that a point on this PR line, either with the original negative slope or the changed positive one? Well, for the P-Zero scores, the mean of them is zero since they all start at zero, and in any individual game, the same quantity is added and subtracted from the winner and loser, respectively. So, $\bar{x}$ for this sample is zero. If $\bar{x}$ is zero, then |m|*zero is zero. And, $\bar{y}$ is the mean of the sample of players that represent our tournament is the only thing left on that side. Thus, the most important point, the centroid, is preserved. So, the absolute value bars on the slope works very well in this situation as it preserves the centroid, and it allows differentiated performance ratings in the correct direction in the face of a negatively sloped regression line.

We shall now explore a few examples. The first is a tournament won by Judit Polgar. I am a big fan of hers. I don't get to play over very many of her games, but the ones that I have went over are impressive. Anyway, this paper isn't about that. So, let's get to this tournament. It was the Isle of Lewis, 1995 tournament. The data as I input it into my app is as follows:

Isle_of_Lewis_1995
4
Chessmetrics(FailedHere)
2687 Polgar
2607 Agdestein
2499 Motwani
2707 Short
2807
3 2 0.5 1 4 1.0
2 4 1.0 3 1 0.0
1 2 0.5 4 3 1.0
2 3 0.0 4 1 0.0
2 1 0.5 3 4 0.0
1 3 1.0 4 2 0.0

When I ran the Chessmetrics data for this tournament, Nigel Short had a large residual that contributed too much to the chi-squared test statistic, and the tournament failed the test. Rather than fix it, I tried to see if I could get the Elo ratings data for the same tournament since it was in 1995. Fortunately, I did get the data. It was as follows:

Isle_of_Lewis_1995
4
Elo
2630 Polgar
2600 Agdestein
2510 Motwani
2655 Short
2765

So, running the app with these initial ratings produced the following P-Zero scores as x, and I will set them up with the initial ratings, y, as ordered pairs.

(17.82, 2630)
(4.49, 2600)
(-8.90, 2655)
(-13.40, 2510)

Running a regression analysis on these data points produces the following regression line equation, the value of r, and the centroid.

r = 0.47
Regression Eq: yhat = b0 + b1*x
Regression Eq: yhat = 2598.75 + 2.0970*x (27)
(xbar, ybar) = (-0.00, 2598.75)

So, our performance rating equation for this tournament is the following:

PR = 2.0970*P-Zero + 2598.75 (28)

When we plug in each of the P-Zero scores for the players above, we get that the performance ratings for each of the players is the following in chart form.

Rank	Name	P-Zero Rating	(O)	PR (E)	Residual (O – E)
1)	Polgar	17.82	2630	2636.12	-6.12
2)	Agdestein	4.49	2600	2608.16	-8.16
3)	Short	-8.90	2655	2580.08	74.92
4)	Motwani	-13.40	2510	2570.65	-60.65

Observe that the residuals are already part of the computation needed for the test statistic. All we need to do is to square the residuals and add those up. Then, just divide by 2500. Doing this produces a test statistic of chi-squared = 3.758. Comparing this against the critical value of chi-squared = 7.810 (the usual chi-squared critical value for alpha = 0.05 for n = 4 implying a d.f. = 3), we see that the tournament has passed. So, we may proceed with the rest of the analysis regarding the G-scores and tournament strength. Judit scored a 203.15 on the greatness scale, which is amazing. In another tournament of hers (Madrid 1994), she scored 203.61, so this is two of her performances breaking into the 200 level. The Isle of Lewis tournament scored a 80.0321 for its strength.

The next tournament we shall discuss is the U.S. Open 2013 officially won by Gata Kamsky. However, the P-Zero score for Alejandro Ramirez was just a bit higher, which means that he really outperformed Kamsky by a tiny amount and deserves to have won the championship. I wrote about this tournament already in a case study, so I won't repeat that here. However, I will say that in that paper I said that we could just round off the quantum correction to a convenient value. Now, I am amending that practice to being rounding it off to the nearest whole number. It doesn't make that much of a difference, really. But, I am just stating it here for the purposes of precision. [Actually, it can make a big difference as the follow up to this paper shows.] The paper on the 2013 US Open and its follow up (Ans, I also posted it in the Greatness of Chess Archive) that I wrote is on Google Drive at the following web address:

https://drive.google.com/file/d/0By2fC4qQL9sKODNnSUNWWkgwRTA

The next tournament we shall discuss is the Hastings 1919 tournament. Hastings has been an elite level tournament for many years. In 1921, Boris Kostic won it with a perfect score. In 1919, Capablanca won it with a nearly perfect score. His only draw was with Boris Kostic, so he played a spoiler role for Capablanca here. As another piece of interesting chess history. Olland got second place in Amsterdam 1899 to the perfect score winner Henry Atkins. Olland got to see history in that tournament. He almost got to see it 20 years later in another tournament he participated in, which is the one we are discussing now—Hastings 1919. I didn't put this tournament in the archive because I will still think about it. But, my analysis for it up until now I will present in this paper. The systemic

mistake that Chessmetrics is making is present in this tournament as well. Kostic has a higher performance rating on the Chessmetrics site for this tournament than Capablanca does because the only two ratings present are these two players, and they drew each other. So, there are just four performance ratings shown on the Chessmetrics site, and Capablanca's is below three other players (two of them are tied but below Kostic, who has the highest). The performance ratings need to account for how the players did against all of the players in the tournament, not just the ones with ratings. Thus, I am correct in using the Ratings Goodness of Fit Test to get the remaining ratings, so that the initial ratings can be regressed on the P-Zero scores. This is an extremely important step. So, I will explain how I have processed this tournament so far.

There is very little ratings information about this tournament. Normally, my go to place for historical tournaments, Chessmetrics, didn't quite have the goods for this tournament. So, I need to fill in the details with the ratings tools I have developed, and the Ratings Goodness of Fit Test plays an important role. Capablanca, in this tournament was rated 2871, which is a very high rating—even for the number one player for Chessmetrics. I need the number two player's rating in a particular rating system to calculate the G-score, which is Lasker at 2848—again a very high rating. We can keep these ratings if they pass the ratings test along with everyone else's ratings. So, here is the original player roster with the Chessmetrics ratings.

2671 Kostic
2545 Thomas
2485 Wahltuch
2871 Capablanca
2497 Yates
2563 Michell
2480 Olland
2468 Scott
2468 Marchand
2500 Conde
2563 Winter
2500 Cole
2846

Now, I have to clarify that these actually are not the Chessmetrics ratings. These are the Chessmetrics ratings for time periods as close to this tournament as is possible for all the players except for Kostic and Capablanca. Kostic and Capablanca have Chessmetrics ratings for this tournament. The rest of the players do not. So, I laboriously clicked on each player and went through their ratings information history on the Chessmetrics site. Chessmetrics allows one to do this. So, I filled in the rest of the ratings information. Lasker is on the ratings list for the month of this tournament, so his rating is official—at least in Chessmetrics terms. However, running this tournament with these ratings failed the Ratings Goodness of Fit Test. Here is a table of the residuals and other relevant information.

Rank)	Name	P-Zero	Perf. Rating	Init. Rating	Residual
1)	Capablanca	44.29	2718.37	2871	152.63**LR**
2)	Kostic	35.40	2684.76	2671	-13.76
3)	Yates	13.44	2601.73	2497	-104.73**LR**
4)	Thomas	13.21	2600.86	2545	-55.86
5)	Michell	-0.04	2550.76	2563	12.24
6)	Wahltuch	-0.15	2550.36	2485	-65.36
7)	Scott	-8.80	2517.63	2468	-49.63

8)	Olland	-8.87	2517.38	2480	-37.38
9)	Marchand	-13.22	2500.95	2468	-32.95
10)	Conde	-17.75	2483.81	2500	16.19
11)	Winter	-22.09	2467.41	2563	95.59
12)	Cole	-35.43	2416.97	2500	83.03

Note: **LR** means Large Residual when its magnitude exceeds 100.

If I fix Yates at 2602 (his performance rating rounded to the nearest whole number), the tournament still fails the Ratings Goodness of Fit test when I run it again. So, I don't want to bring Capablanca down to 2718 to fix his rating because that is too much for the number one player to fall, especially since Lasker is at 2846. I would have to lower him too since the number one can't be below the number two especially at more than one hundred points lower. So, this is a real conflict here. How can we resolve this situation? Well, we are going to have to break one of my protocols.

After some deliberation, I will resolve this by dropping Winter by 75 points and lowering Capablanca and Lasker each by 50 points. This goes against my rules that I have laid out, but this situation is for a tournament about 100 years old now. There is enough lack of information on these players to enable me to bend my rules slightly. This bending of the rules won't be necessary with current and accurate rating information on the players in tournaments of today. It will work just as I have shown. So, bending the rules slightly to accommodate this situation yields the following initial rating list.

2671 Kostic
2545 Thomas
2485 Wahltuch
2821 Capablanca
2602 Yates
2563 Michell
2480 Olland
2468 Scott
2468 Marchand
2500 Conde
2488 Winter
2500 Cole
2796

Observe Lasker (not named, but the 2796 rating is his representing the rating of the #2 player in the population) has been lowered by 50 points as stated. Running the app with these initial ratings produces the following analysis related to the regression line and the Ratings Goodness of Fit Test.

Regression Analysis:

r = 0.87
Regression Eq: yhat = b0 + b1*x
Regression Eq: yhat = 2549.25 + 3.9239*x (29)
(xbar, ybar) = (0.00, 2549.25)

Residual Analysis (Magnitudes > 100 Become Suspect):

Rank)	Name	P-Zero	Perf. Rating	Init. Rating	Residual
1)	Capablanca	44.29	2723.05	2821	97.95
2)	Kostic	35.40	2688.17	2671	-17.17

3)	Yates	13.44	2601.99	2602	0.01
4)	Thomas	13.21	2601.09	2545	-56.09
5)	Michell	-0.04	2549.09	2563	13.91
6)	Wahltuch	-0.15	2548.67	2485	-63.67
7)	Scott	-8.80	2514.70	2468	-46.70
8)	Olland	-8.87	2514.45	2480	-34.45
9)	Marchand	-13.22	2497.39	2468	-29.39
10)	Conde	-17.75	2479.61	2500	20.39
11)	Winter	-22.09	2462.58	2488	25.42
12)	Cole	-35.43	2410.23	2500	89.77

Note: **LR** means Large Residual when its magnitude exceeds 100.

Next, is a better test of the goodness of fit of the regression line using chi-squared.
For my research, I am keeping a list of C.I.'s for E, my constant in the
test statistic computation. E is in (1556.6, 6696.2).
H0: The players all performed to expectation.
HA: The players did not all perform to expectation.
Critical Value: chi-squared = 19.680.
Lower Threshold Value: chi-squared = 4.575.

If the test statistic exceeds the critical value, that is evidence that the players did
not all perform according to expectation (the performance rating as calculated by the
regression line equation). Goodness of fit tests are all right tailed, so we just need
to make this one comparison. Test Statistic: chi-squared = 12.254.
The test statistic is in the do not reject region. We conclude that the players
performed as expected.

Hastings_1919

Rank	Name	Chessmetrics (PRA-corrected)	Elo After	PRA After	Perf. Rating
1)	Capablanca	2821	2832.90	2840.81	2723.05
2)	Kostic	2671	2690.08	2696.74	2688.17
3)	Yates	2602	2607.85	2613.34	2601.99
4)	Thomas	2545	2558.76	2559.57	2601.09
5)	Michell	2563	2559.90	2563.35	2549.09
6)	Wahltuch	2485	2493.50	2489.59	2548.67
7)	Scott	2468	2470.72	2464.90	2514.70
8)	Olland	2480	2480.50	2476.25	2514.45
9)	Marchand	2468	2465.98	2460.62	2497.39
10)	Conde	2500	2487.93	2486.27	2479.61
11)	Winter	2488	2473.51	2470.80	2462.58
12)	Cole	2500	2469.37	2468.77	2410.23

The Ratings Goodness of Fit Test passed. This means that the results
of the tournament fit with the historical ratings. No quantum corrections
are necessary. The results are hereby finalized and certified.

Note that now there are no large residuals. Capablanca and Cole have the worst ones, but since both are less than 100 points out, they are in spec; however, both of these players have what I now call downward pressure on their ratings. It is OK for a tournament to have one large residual (as long as it isn't too large) because such a tournament can still pass the Ratings Goodness of Fit Test. Also, it may be possible for all the ratings to be in spec and still fail this test. If all are from, say, 90

to 99.9 points in magnitude, that would probably do it (large upward and/or downward pressures). But, the tournament now passes the test. Because it now passes, we can report the G-scores and tournament strength.

Greatness Scores:
Capablanca (Hastings_1919) got G = 285.81
Kostic (Hastings_1919) got G = 233.36
Hastings_1919: Tournament Strength: 71.3432

So, Capablanca didn't do quite as well as his New York Rice 1913 tournament where he scored over 300 on the G-scale. But, make no mistake about it, a 285 is very good indeed. I just finished processing Kasparov's Belgrade 1989 tournament where he actually scored a higher raw percentage than his supposed greatest tournament Tilburg 1989. In terms of points, Kasparov got 9.5 out of 11 while Capablanca got 11.5 out of 12. As a raw percent, Capablanca obviously did better, but when we account for the tournament strength, it isn't so clear then. But, the G-score is the computation that takes into account everything. So, Capablanca at Hastings 1919 gets a G-score of 285.81 in a tournament with a strength of 71.3432. Kasparov's Belgrade 1989 G-score was 280.11 with a tournament strength of 85.4344. So, by using the G-score comparison, we can say that Capablanca here did better than Kasparov did in Belgrade 1989. We can only say that because the Ratings Goodness of Fit Test enables us to calibrate the ratings of the Hastings 1919 tournament, which then enables us to compute an accurate G-score for Capablanca at that tournament.

Hastings 1895 is the next tournament I would like to discuss. Harry Nelson Pillsbury won this tournament before he was recognized as a great player. There is a great deal of information about this tournament online. Indeed, it is the most famous tournament of the 19th century. Why do I select this one to discuss? It is because Chessmetrics has two ratings out of spec, and one missing rating, and we can discover how to fix the out of spec ones and fill in the missing one using the Ratings Goodness of Fit Test. Pollock was the other player out of spec as mentioned earlier in this paper. I will assume the reader has reviewed that tournament from the archive, so I don't need to repost that information again here.

The protocol I have established in the archive for missing ratings is to pair up the initial ratings you do have with the P-Zero scores. Since the P-Zero scores are more complete than the initial ratings, we can use them to fill in what the initial ratings might have been. However, the data pairs have to be regressed y on x (or the initial ratings on the P-Zero scores). Then, test this regression line with the Ratings Goodness of Fit Test. If it passes, then it is ok to use the line to fill in the remaining ratings by computing y(P-Zero) for each player with a missing rating. If the line fails the test, the delete the worst out of spec (or the worst contributor to chi-squared), and put this player in the pool of players whose rating is unknown. Repeat this process until you get a data set that passes the test. I have not needed to throw out three initial ratings from Chessmetrics. But, I have had to throw out two. Hastings 1895 and Netanya 1968 are two examples of cases I had to throw out two Chessmetrics ratings. Perusing the archive shows a colorful mixture of missing and incorrect ratings from Chessmetrics—all fixed by my system. When missing ratings are supplied, I call the ratings the XXXX(PRA-Corrected) ratings where XXXX is the name of the rating system used.

If you have all of the initial ratings, but it fails the Ratings Goodness of Fit Test, then the standard protocol to follow is to move the out of spec ratings to the nearest whole number on the regression line. Then, rerun the tournament in the app with these fixed ratings. However, other protocols might be followed. It depends on the real world situation. The tournament app does not know what is going on in the real world. An example I have where I did some rather drastic things is Cambridge Springs 1904, Marshall's greatest tournament victory. But, I must say this tournament passes the Ratings Goodness of Fit Test because of the placement of the errors was just perfect. I explain this

in the archive for this tournament. I observed some real world extremely relevant information that changed several of the initial ratings. Lasker's rating was set to his London 1899 rating without the Chessmetrics decay feature ruining Lasker's rating because he hadn't played for five years. Clearly, Lasker played to this rating. Maybe he had a little rust, but that was all knocked off by the end of the tournament when he was in third with one point behind the second place Janowski. Lasker's defeat of him actually garnered him second place, according to the performance ratings for this tournament. Marshall won this tournament convincingly with a two point margin of victory over Lasker. Pillsbury's rating had to be lowered because he had syphillus, and it was at a stage of the disease where Pillsbury decided that was his last tournament. One of the truly great players (a G-score of 373.57, second best of all time as of this writing), it is very probable he could have taken Lasker's crown before 1921 when Capablanca won it. These adjustments were made by setting Lasker's rating, and deleting Marshall's and Pillsbury's ratings to be determined by the system. The final set of intial ratings garnered also pass the Ratings Goodness of Fit Test, but they capture the real world situation better. When just adjustments are made with all of the initial ratings present, I call the new ratings the XXXX(PRA-Adjusted ratings where the XXXX is the name of the rating system.

Upon failure of the test, one might do nothing. The automatic nature of the rating system will fix the ratings over time if the Ratings Administrator gets it wrong. However, the Ratings Administrator is there to attempt to get it right. I suppose some tracking of how an administrator is doing would be in order. If one is getting it wrong too many times to "help boost the egos of his buddies," then some disciplinary action might be in order. But, the only master of the administrator is the integrity of the ratings, so if each one really believes in that, errors should be few and far between. I did nothing on several tournaments. The US Championship 2015, I did nothing because a number of other players showed pressure on their ratings, which also rack up some chi-squared points leading to failure of the tournament. Wesley So was out of spec, but not by much. Sevian was very close to being out of spec, but he had some good upward pressure on his rating. I expect he will climb to over 2600 soon. Troff and Holt should also climb over 2600, but the evidence for them doing so is less convincing than that for Sevian. I also did nothing with the Cap D'Agle 2013 tournament.

If all of the initial ratings are present, and no real world situation caused the failure of the tournament, then just move any or all of the out of spec ratings to the regression line (to the nearest whole number). A ratings administrator gets one and only one opportunity to fix the situation, and then the automatic nature of the rating system takes over. Rerun the app with the new initial ratings intact. If the tournament still fails, it will be by less. However, in my experience building the archive, fixing just one out of spec (when all of them are there initially), will allow the tournament to pass. There were three out of specs in the US Championship (Sevian, Troff, and Kamsky). Fixing two of them in the first part of that case study (included in this document) allowed the tournament to pass. However, in the second part of that case study, I showed how moving all three of these players to the regression line rounded to the nearest whole number produced the most accurate set of final ratings for the players at the conclusion of the tournament. So, the moral of this example is that the rating system sees things that should be taken into account. Only if there is a compelling reason, the ratings administrator should follow the recommendations of the system, and adjust all of the out of specs. As a reminder, I am still investigating the upward pressure and downward pressure issue. I have a fine example in the 2015 US Championship that just concluded with one barely out of spec rating (Wesley So) and four ratings with pressure (Sevian, Troff, and Holt had upward pressure, and Naroditsky had downward pressure). Over the next several months, I will be watching the 12 players of this tournament to see if doing TRA's (Tournament Rating Adjustments) on players with pressure on their ratings is useful or not.

Of course, if the tournament passes, then nothing needs to be done. Just let the system run its course.This happened a number of places in the archive. Palma de Mallorca 1970 was a

Chessmetrics example that went without a hitch. Karpov's 1994 Linares tournament is an Elo example that went without a hitch. Most of the Elo tournaments actually have passed. I have found very little to quarrel about with the Elo system compared to the Chessmetrics system. But, Elo is clearly not free from a number of problems as I have pointed out. Chessmetrics has more problems than Elo as I see it. Sonas claims his system is unsurpassed, yet FIDE still uses Elo and not his Chessmetrics system. They have their reasons unbenownst to me, but based on my findings, they are correct in not coming off of Elo in favor of Chessmetrics. However, they would be wise to adopt the PRA and Tournament App system because that is a vast improvement, and the tie-breaking abilities of the system are unsurpassed (to use Sonas's word).

Observe in the last column of the final chart provided in any tournament in the archive that the performance ratings now descend beautifully down the chart. This is a fine indicator of correctness. If you look at current FIDE charts of performance ratings (the TPR's), they do not exhibit this behavior. Thus, I have ascertained the true performances with these equations and procedures I have developed and explained here. Once the administrator does whatever is going to be done based on the results of the Ratings Goodness of Fit Test, he or she can report any G-scores as well as the strength of the tournament.

From the tournaments I have processed using my algorithm and procedures, I have a few nearby performances to Pillsbury's and Chigorin's performances at Hastings 1895. I will just copy and paste the 240's and 250's from my Greatness List. Note that a G-score of 250 is a convenient cutoff for what I am calling a legitimate threat to the championship. Pillsbury at Hastings 1895 is seen here to be surrounded by excellent company. Fischer became champion in a few years after his performances in Sousse 1967 and Netanya 1968. True enough, he didn't finish the tournament in Sousse, but he did enough to get a G-score higher than Larsen. For a while Carlsen's Pearl Springs 2009 performance was bantied about as being superior to Karpov's Linares 1994 performance, but we see here that his G-score was substantially lower than Karpov's G-score. Nevertheless, Carlsen did score above the 250 mark in that tournament, and obviously, he went on to become the champion. Steinitz got second place in London 1883 to Zuckertort, but it was still over 250. We know he defeated Zuckertort for the title. He was losing to Zuckertort at first, but then he implemented his new system, which exposed Zuckertort's main weakness of not having an overall plan. Some other incredible performances are here in this small list: Sofia Polgar's Rome 1989 had a very high TPR (a faulty calculation, but it induced me to include her here), and Sargissian's 2007 performance also had a high TPR (I included him for the same reason), and Gligoric, Kostic, and Maroczy are here, too. Kasparov at Linares 1993 scored a G = 244.06, which is just a bit shy of the 250 mark; but I have observed that even a champion will score above 250 every time. In fact, the champion should only be expected to score above 250 maybe only a couple of times during his or her tenure as champion. Kostic was undefeated at Hastings 1921, but the opposition was less that what Kasparov faced in 1993 in Linares. I need to fill out this list a lot more. Many great players have performances that I simply have not gotten to yet. I need to write something that can process my stuff on the million games plus databases so that I can do major blocks of chess history at one shot. The ones that fail the Ratings Goodness of Fit Test can be inspected personally should they occur.

66) Carlsen (Pearl_Springs_2009) got G = 255.46
67) Polgar_S (Rome_1989) got G = 255.15
68) Steinitz (London_1883) got G = 253.55
69) Pillsbury (Hastings_1895) got G = 253.16
70) Kostic (Hastings_1921) got G = 252.66
71) Fischer (Sousse_Interzonal_1967) got G = 251.06
72) Sargissian (Ruy_Lopez_Festival_2007) got G = 248.99
73) Chigorin (Hastings_1895) got G = 246.66

74) Gligoric (Reykjavik_1964) got G = 246.60
75) Maroczy (Monte_Carlo_1903) got G = 244.40
76) Kasparov (Linares_1993) got G = 244.06
77) Fischer (Netanya_1968) got G = 240.58

Here are the tournaments scoring in the 79's, 80's, and 81's, and among them is the Hastings 1895 tournament. From this list, we can see that these tournaments were all about as strong as the Hastings 1895 tournament. I have taken the resistance coefficient from the Greatness Formula, and computed a weighted average of it for each player based on the number of games each player contributed to the tournament. This method provides a very sharp method of ranking the tournaments of history because of the incredible refinement of this coefficient. The Greatness Formula will be discussed in the next chapter. The paper will conclude with a few case studies and a couple of other essays.

76) Batumi_2001: 81.8353
77) USSR_Championship_1964: 81.5727
78) Dresden_1926: 81.4174
79) US_Championship_2015: 81.1288
80) US_Open_1964: 80.7531
81) Hastings_1895: 80.3996
82) Monte_Carlo_1903: 80.3941
83) Ruy_Lopez_Festival_2007: 80.1713
84) Stockholm_Interzonal_1962: 80.1038
85) Isle_of_Lewis_1995: 80.0321
86) San_Remo_1930: 79.1783
87) Riga_Interzonal_1979:79.1435

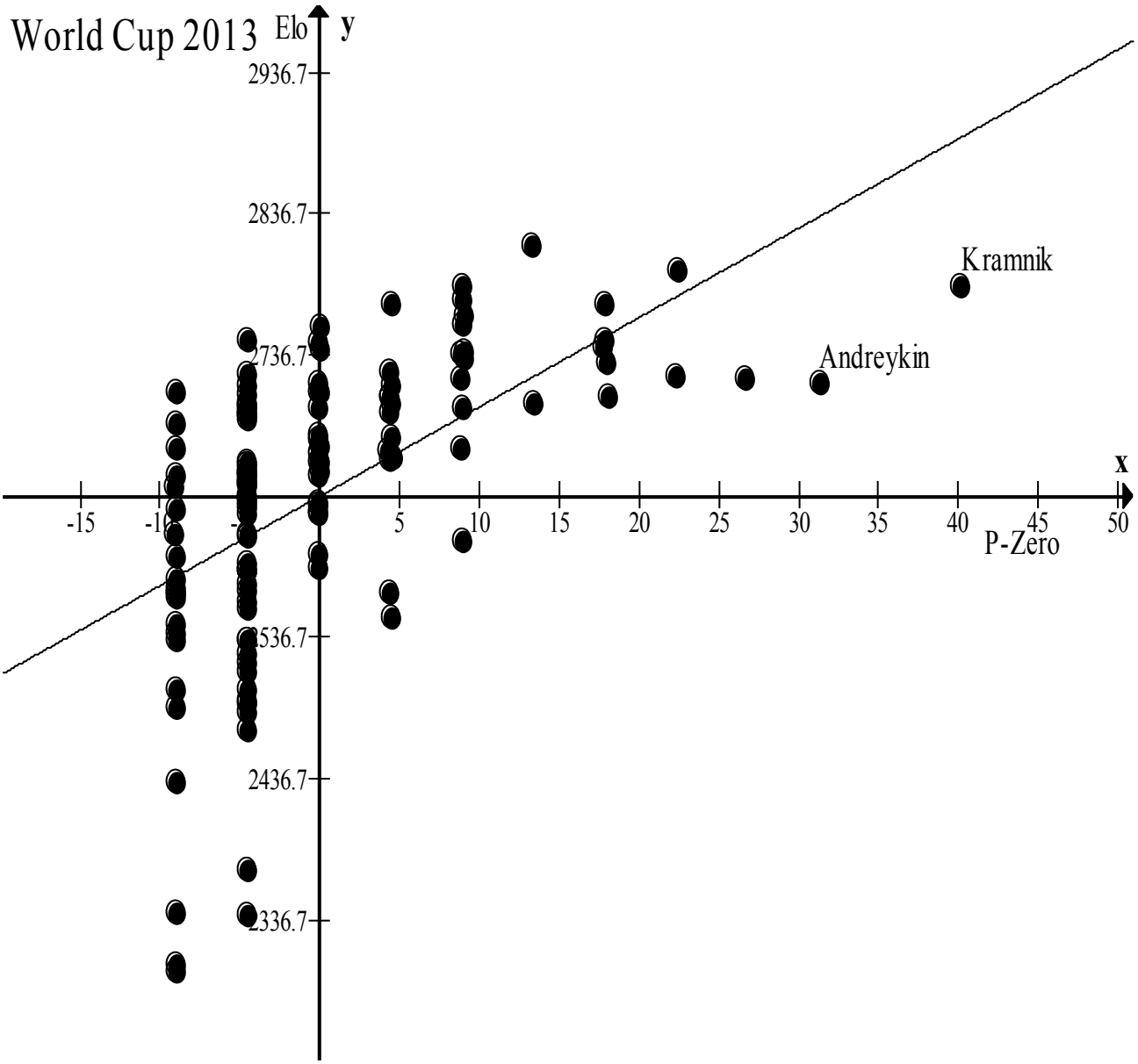
As an important observation, the 1964 US Open was the tournament that Bobby Fischer routed 11-0. It has been criticized as not being very strong, so Fischer has taken a bit of a beating in chess circles. Note that it just edges out Hastings 1895 in terms of its strength. They claimed the all grandmaster field in the 2015 version of the tournament was the strongest in history, but we can see here that it was less than half a point stronger than the version Fischer participated in. Would he have routed the competition in the 2015 edition? Well, it's debatable for sure, but it certainly was basically the same strength as the 1964 edition, which he did route. My equations are capable of shedding light on these issues that were not really resolved through debate before. I didn't notice that these two tournaments (the 2015 and 1964 versions of the US Open/Championship) were that close in strength until I pasted that fragment of the tournament list here. The other tournaments listed here in this fragment are certainly noteworthy as well.

Next, we will discuss the World Cup 2013. This one is here to show how the knock out style of tournament is a failure for rating systems. The calculation of a rating is possible only over the course of a number of games. The knock-out style of tournament is a punishing feature that causes higher rated players to be “knocked out” of the tournament early on. The lower rated players get to move on and play more games. In this particular tournament under discussion, there were 128 players, so I am not giving full details. In fact, I will just show the graph and report the Ratings Goodness of Fit Test results. The graph is on the next page. The lowest rating was about 2300, and the highest was 2800. There were quite a number of chess greats at this tournament. It was won by Vladimir Kramnik, a former world champion.

Residual Analysis (Magnitudes > 100 Become Suspect):

Rank) Name: P-Zero, Perf. Rating, Init. Rating, Residual

1) Kramnik97: 40.17, 2890.52, 2784, -106.52**LR**
2) Andreykin53: 31.36, 2834.85, 2716, -118.85**LR**
3) Lagrave95: 26.76, 2805.76, 2719, -86.76
4) Karuana66: 22.44, 2778.50, 2796, 17.50
5) Korobov128: 22.36, 2778.00, 2720, -58.00
6) Tomashevsky4: 18.13, 2751.28, 2706, -45.28
7) Ivanchuk111: 18.00, 2750.43, 2731, -19.43
8) Svidler35: 17.95, 2750.12, 2746, -4.12
9) Nakamura114: 17.89, 2749.74, 2772, 22.26
10) Kamsky16: 17.81, 2749.24, 2741, -8.24
11) Liem33: 13.42, 2721.48, 2702, -19.48
12) Aronian1: 13.36, 2721.15, 2813, 91.85
13) Gelfand81: 9.11, 2694.26, 2764, 69.74
14) Radjabov46: 9.08, 2694.08, 2733, 38.92
15) Morozevich10: 9.08, 2694.06, 2739, 44.94



World Cup 2013

16) Mamedyarov19: 9.00, 2693.58, 2775, 81.42

17) Grischuk32: 8.97, 2693.41, 2785, 91.59
18) Dominges86: 8.97, 2693.40, 2757, 63.60
19) Hammer25: 8.97, 2693.40, 2605, -88.40
20) Moiseenko91: 8.95, 2693.23, 2699, 5.77
21) Dubov125: 8.94, 2693.17, 2669, -24.17
22) Hiro80: 8.89, 2692.89, 2737, 44.11
23) Vityugov13: 8.84, 2692.56, 2719, 26.44
24) Granda78: 4.61, 2665.83, 2664, -1.83
25) Karyakin47: 4.56, 2665.51, 2772, 106.49**LR**
26) Yi29: 4.55, 2665.49, 2551, -114.49**LR**
27) Bakro84: 4.50, 2665.15, 2714, 48.85
28) Krivoruchko109: 4.50, 2665.15, 2678, 12.85
29) Elyanov59: 4.50, 2665.14, 2702, 36.86
30) Yu73: 4.47, 2664.98, 2662, -2.98
31) Vallejo42: 4.47, 2664.97, 2706, 41.03
32) Shirov37: 4.47, 2664.97, 2696, 31.03
33) Adhiban116: 4.45, 2664.81, 2567, -97.81
34) Jakovenko50: 4.45, 2664.80, 2724, 59.20
35) Malakhov67: 4.42, 2664.63, 2707, 42.37
36) Dreev61: 4.37, 2664.30, 2668, 3.70
37) Ponomariov117: 0.09, 2637.30, 2756, 118.70**LR**
38) Sverch40: 0.08, 2637.23, 2654, 16.77
39) Safarli121: 0.08, 2637.23, 2660, 22.77
40) Adams102: 0.05, 2637.06, 2740, 102.94**LR**
41) Areschenko100: 0.05, 2637.06, 2709, 71.94
42) Bologan44: 0.05, 2637.06, 2672, 34.94
43) Onischuk93: 0.03, 2636.89, 2667, 30.11
44) Sashikiran58: 0.03, 2636.89, 2660, 23.11
45) Ragger22: 0.03, 2636.89, 2680, 43.11
46) Kobaliya105: 0.03, 2636.89, 2651, 14.11
47) Leytao12: 0.00, 2636.72, 2632, -4.72
48) Leko69: 0.00, 2636.72, 2744, 107.28**LR**
49) Dzhobava63: 0.00, 2636.72, 2716, 79.28
50) So7: 0.00, 2636.72, 2710, 73.28
51) Filippov89: -0.00, 2636.72, 2630, -6.72
52) Matlakov28: -0.00, 2636.72, 2676, 39.28
53) Fier123: -0.00, 2636.72, 2595, -41.72
54) Robson119: -0.03, 2636.55, 2623, -13.55
55) Trung_Son71: -0.03, 2636.55, 2625, -11.55
56) Felgaer108: -0.03, 2636.55, 2586, -50.55
57) Bruson55: -0.08, 2636.21, 2698, 61.79
58) Lysy6: -4.45, 2608.63, 2648, 39.37
59) Shimanov23: -4.45, 2608.63, 2655, 46.37
60) Suarez103: -4.47, 2608.46, 2609, 0.54
61) Iturrizaga94: -4.47, 2608.46, 2660, 51.54
62) Darini54: -4.47, 2608.46, 2535, -73.46
63) Yohannessen70: -4.47, 2608.46, 2519, -89.46
64) Kaidanov99: -4.47, 2608.46, 2574, -34.46
65) Shoker20: -4.47, 2608.46, 2489, -119.46**LR**
66) Akopyan72: -4.47, 2608.46, 2691, 82.54
67) Ipin15: -4.47, 2608.46, 2484, -124.46**LR**
68) Yu.Polgar104: -4.47, 2608.46, 2696, 87.54

69) Kristiansen75: -4.47, 2608.46, 2584, -24.46
70) Hismatullin106: -4.47, 2608.46, 2653, 44.54
71) Ryazantsev107: -4.47, 2608.46, 2700, 91.54
72) Ipatov8: -4.47, 2608.46, 2584, -24.46
73) Salem79: -4.47, 2608.46, 2556, -52.46
74) Mareko18: -4.47, 2608.46, 2561, -47.46
75) Azarov62: -4.47, 2608.46, 2636, 27.54
76) Rahman82: -4.47, 2608.46, 2470, -138.46**LR**
77) Sebbar48: -4.47, 2608.46, 2371, -237.46**LR**
78) Hansen118: -4.47, 2608.46, 2492, -116.46**LR**
79) Akash65: -4.47, 2608.46, 2340, -268.46**LR**
80) Postny88: -4.47, 2608.46, 2628, 19.54
81) Movsesyan26: -4.47, 2608.46, 2699, 90.54
82) Voytashek124: -4.47, 2608.46, 2701, 92.54
83) Istratesku5: -4.47, 2608.46, 2646, 37.54
84) Hansen68: -4.47, 2608.46, 2584, -24.46
85) Wang_Hao52: -4.47, 2608.46, 2747, 138.54**LR**
86) Inarkiev11: -4.47, 2608.46, 2693, 84.54
87) Sambuev9: -4.50, 2608.29, 2524, -84.29
88) Kori45: -4.50, 2608.29, 2569, -39.29
89) Popov21: -4.50, 2608.29, 2644, 35.71
90) Nepomnyaschy30: -4.50, 2608.29, 2723, 114.71**LR**
91) Smits27: -4.50, 2608.29, 2623, 14.71
92) Duda112: -4.50, 2608.29, 2534, -74.29
93) Navara17: -4.50, 2608.29, 2715, 106.71**LR**
94) Hrachek43: -4.50, 2608.29, 2635, 26.71
95) Yungo101: -4.50, 2608.29, 2512, -96.29
96) Chao87: -4.50, 2608.29, 2693, 84.71
97) Lupulesku57: -4.50, 2608.29, 2634, 25.71
98) Fressine76: -4.50, 2608.29, 2708, 99.71
99) Romanov90: -4.50, 2608.29, 2651, 42.71
100) Dzhons24: -4.50, 2608.29, 2645, 36.71
101) Tsinnan51: -4.50, 2608.29, 2500, -108.29**LR**
102) Brunello60: -4.50, 2608.29, 2607, -1.29
103) Ramires3: -4.50, 2608.29, 2588, -20.29
104) Bassem122: -4.53, 2608.12, 2652, 43.88
105) Zvyagintsev39: -4.53, 2608.12, 2659, 50.88
106) Melkumyan77: -4.53, 2608.12, 2632, 23.88
107) Ushenina36: -8.95, 2580.20, 2500, -80.20
108) Kravtsiv64: -8.95, 2580.20, 2535, -45.20
109) Alekseev115: -8.95, 2580.20, 2710, 129.80**LR**
110) Bvaliya98*: -8.97, 2580.03, 2300, -280.03**LR**
111) Kori113: -8.97, 2580.03, 2434, -146.03**LR**
112) Holt14: -8.97, 2580.03, 2539, -41.03
113) Paragua49: -8.97, 2580.03, 2565, -15.03
114) Naer56: -8.97, 2580.03, 2626, 45.97
115) Barbosa34: -8.97, 2580.03, 2571, -9.03
116) Belobrk31: -8.97, 2580.03, 2341, -239.03**LR**
117) Flores41: -8.97, 2580.03, 2578, -2.03
118) Volokitin120: -8.97, 2580.03, 2688, 107.97**LR**
119) Adly92: -8.97, 2580.03, 2594, 13.97
120) Belyavsky74: -8.97, 2580.03, 2651, 70.97

121) Agdestein83: -8.97, 2580.03, 2567, -13.03
122) Markov2: -8.97, 2580.03, 2304, -276.03**LR**
123) Shabalov96: -8.97, 2580.03, 2546, -34.03
124) Fedorchuk126: -8.97, 2580.03, 2669, 88.97
125) Durarbeyli127: -8.97, 2580.03, 2567, -13.03
126) Gindi85: -8.97, 2580.03, 2487, -93.03
127) Yifan38: -9.00, 2579.86, 2609, 29.14
128) Negi110: -9.03, 2579.69, 2643, 63.31
Note: **LR** means Large Residual when its magnitude exceeds 100.

Next, is a better test of the goodness of fit of the regression line using chi-squared.
For my research, I am keeping a list of C.I.'s for E, my constant in the test statistic computation. E is in (5712.2, 8643.6).

H0: The players all performed to expectation.
HA: The players did not all perform to expectation.

Critical Value: chi-squared = 154.300.
Lower Threshold Value: chi-squared = 101.971.

If the test statistic exceeds the critical value, that is evidence that the players did not all perform according to expectation (the performance rating as calculated by the regression line equation). Goodness of fit tests are all right tailed, so we just need to make this one comparison.

Test Statistic: chi-squared = 352.556.
The test statistic exceeds the critical value. The players did not perform as expected.

So, it was a pretty spectacular failing reported by the Ratings Goodness of Fit Test. I had to spend a lot of time getting all of the information ready to feed into my app. The partial results that I am displaying show quite a number of the residuals as being out of spec. I will go on record here as saying that the failings displayed are not the fault of the players. Highly rated players who got "knocked out" early did not have the opportunity to show their greatness because they were removed from the tournament. Lower rated players who got lucky and knocked out a higher rated player got to move on. These kinds of things can happen in the knock-out style of tournament. When they do, a higher rated player's performance rating will not coincide with his (or her) actual rating. So, I have decided to do nothing with this tournament as far as quantum corrections are concerned. No ratings will be fixed (i.e. no quantum corrections will take place). So, these ratings will stand, and the greatness scores and tournament strength will be reported without alteration. Therefore, here are the results from this tournament.

Greatness Scores:
Kramnik97 (World_Cup_2013) got G = 212.04
Andreykin53 (World_Cup_2013) got G = 153.55
Lagrave95 (World_Cup_2013) got G = 170.28
Karuana66 (World_Cup_2013) got G = 146.28
Korobov128 (World_Cup_2013) got G = 141.42
Ivanchuk111 (World_Cup_2013) got G = 147.01
Svidler35 (World_Cup_2013) got G = 102.87
Nakamura114 (World_Cup_2013) got G = 156.02
Kamsky16 (World_Cup_2013) got G = 101.38

Aronian1 (World_Cup_2013) got G = 114.38
Dominges86 (World_Cup_2013) got G = 100.92

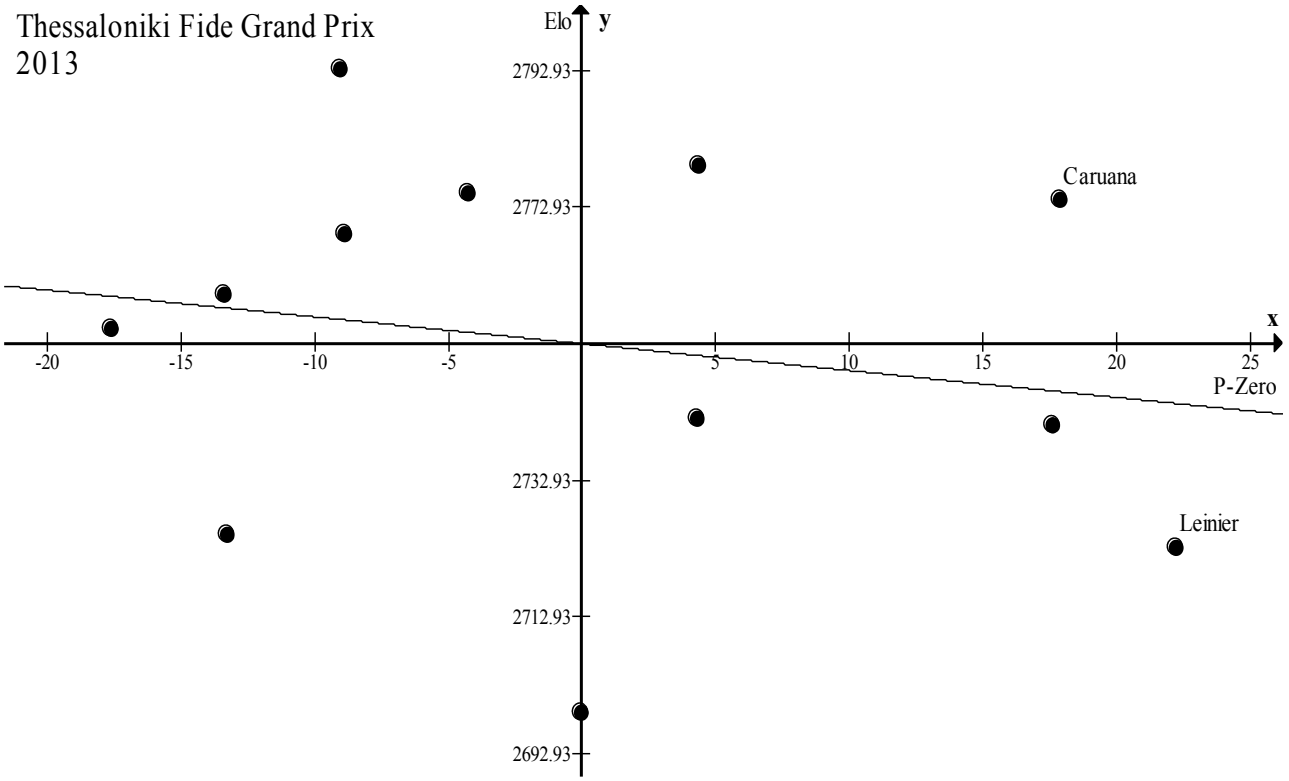
World_Cup_2013: Tournament Strength: 83.7316

So, because the P-Zero scores do not accurately reflect the performances of the players (because of the knock-out format), we should use the Greatness Scores instead. Doing this, we find that Kramnik won the tournament. However, second place goes to Lagrave. And, rounding out the top three, we see that Nakamura deserves third place. Next, I will show a part of a tournament with a negatively sloped regression line.

Thessaloniki Greece Fide Grand Prix 2013

Regression Analysis:

r = -0.17
Regression Eq: yhat = b0 + b1*x
Regression Eq: yhat = 2752.92 + -0.3669*x (30)
(xbar, ybar) = (0.00, 2752.92)



Thessaloniki Fide Grand Prix 2013

Residual Analysis (Magnitudes > 100 Become Suspect):

Rank) Name: P-Zero, Perf. Rating, Init. Rating, Residual
1) Leinier: 22.26, 2744.75, 2723, -21.75
2) Kamsky: 17.65, 2746.44, 2741, -5.44
3) Caruana: 13.49, 2747.97, 2774, 26.03
4) Grishuk: 4.92, 2751.11, 2779, 27.89
5) Ponomariov: 4.66, 2751.21, 2742, -9.21

6) Kasimdshanov: -0.01, 2752.92, 2699, -53.92
7) Nakamura: -4.24, 2754.47, 2775, 20.53
8) Topalov: -4.65, 2754.62, 2793, 38.38
9) Svidler: -8.90, 2756.18, 2769, 12.82
10) Morozevich: -13.42, 2757.84, 2760, 2.16
11) Bacrot: -14.06, 2758.08, 2725, -33.08
12) Ivanchuk: -17.68, 2759.40, 2755, -4.40
Note: **LR** means Large Residual when its magnitude exceeds 100.

Next, is a better test of the goodness of fit of the regression line using chi-squared.
For my research, I am keeping a list of C.I.'s for E, my constant in the
test statistic computation. E is in (412.9, 1776.4).
H0: The players all performed to expectation.
HA: The players did not all perform to expectation.
Critical Value: chi-squared = 19.680.
Lower Threshold Value: chi-squared = 4.575.

If the test statistic exceeds the critical value, that is evidence that the players did
not all perform according to expectation (the performance rating as calculated by the
regression line equation). Goodness of fit tests are all right tailed, so we just need
to make this one comparison. Test Statistic: chi-squared = 3.251.
The test statistic is in the do not reject region. We conclude that the players
performed as expected.

Warning: The test statistic is below the lower threshold value. The players
performed unusually close to expectation as to arouse suspicion.

To compute the Perf. Ratings, we take the absolute value of the slope of the regression equation.
Thus,

yhat = 2752.92 + -0.3669*x (31)

becomes

PR = 2752.92 + 0.3669*P-Zero. (32)

Thessaloniki_Fide_Grand_Prix_2013

Rank	Name	Elo	Elo After	PRA After	Perf. Rating
1)	Leinier	2723	2751.90	2746.38	2761.09
2)	Kamsky	2741	2756.61	2754.64	2759.39
3)	Caruana	2774	2785.92	2786.79	2757.87
4)	Grishuk	2779	2779.65	2782.88	2754.72
5)	Ponomariov	2742	2748.27	2747.06	2754.63
6)	Kasimdshanov	2699	2707.91	2701.18	2752.91
7)	Nakamura	2775	2762.95	2765.51	2751.36
8)	Topalov	2793	2779.57	2786.47	2751.21
10)	Morozevich	2760	2744.19	2746.23	2747.99
11)	Bacrot	2725	2715.14	2712.12	2747.76
12)	Ivanchuk	2755	2736.26	2737.29	2746.43

Note: The slope of the regression line was negative, so the sign of the slope has been made positive
to create ascending performance ratings.

The Ratings Goodness of Fit Test passed. This means that the results of the tournament fit with the historical ratings. No quantum corrections are necessary. The results are hereby finalized and certified.

The foregoing exposition of theory and example tournaments show The Ratings Goodness of Fit Test. It is a solid contribution to our understanding of ratings and the general measurement of performance. Along the way, I have shown a true method of obtaining a performance rating that is not only highly accurate, but also capable of calibrating historical ratings of players through the quantum correction process (also called a TRC, Tournament Rating Correction). Of course, the big question is when do we do this and who do we do it to? The Ratings Goodness of Fit Test answers the when part, and the flagging of the large residuals by the program tells us the who part. It is the judgment of the ratings administrator to decide who among the flagged high residuals as to who actually gets a quantum correction (either up or down). Also, to be decided is the issue of pressure (either upward or downward) and whether or not the administrator should adjust ratings exhibiting pressure.

In order to get the information needed for the Ratings Goodness of Fit Test, the P-Zero scores are necessary. These are obtained from setting all the players to a zero rating and running the PRA in its basic mode only on all the games played in the tournament. The P-Zero scores go on the x-axis, and the ratings of the players go on the y-axis. Each player gets an ordered pair (P-Zero score, rating). We run a regression analysis on these data points. The y-hat from the regression line becomes the E in the Ratings Goodness of Fit Test. The O are the initial ratings of the players. The degrees of freedom are the number of players in tournament minus one. The test statistic is the computation

chi-squared = summation[(O – E)^2/2500]. (33)

We now can look up the critical value and compute our test statistic. If we find that the tournament passes the Ratings Goodness of Fit Test, then the regression line becomes the performance rating equation for this tournament with the caveat that if the slope of the regression line is negative, we take the absolute value of this slope for the slope of the performance rating equation. We can then compute the various performance ratings, and these would be the official ones. If the tournament fails the test, then the ratings administrator may elect to adjust one or several or all of the out of spec ratings flagged by the program. Then, rerun the test with these new ratings. It should pass now. And, with passage of the test, we can compute all the stuff such as G-scores and tournament strength.

So, the PRA has two components: The Basic System, and the Boosting System. If a player receives a quantum correction, then the Boosting System for that player should be turned off. The purpose of the boosting system is to find the ratings of the players when they are getting close to their actual ratings. A quantum correction does this job, so the Boosting System for that player is no longer needed. But, a player may not ever actually participate in a tournament, and so the Boosting System for that player will run its course in time. One other random thought: A player should have played at least 6 games in a tournament to receive a quantum correction. If a player leaves a tournament early for whatever reason, and at least 6 games have not been played by that player, the ratings administrator should not punish that player by lowering his or her rating with a quantum correction.

That concludes the essay on the Ratings Goodness of Fit Test. Next, we will present the essay on the Greatness Formula, and the Tournament Strength formula.

The Greatness Formula

The Greatness Formula is a new formula I devised to measure the greatness of a player (chess or any other competitive modality) during a tournament or match. Given the initial ratings of the players of a tournament (or match), we can plug the match ups along with the raw scores (how the point was split) into the app to get a G-score, where G stands for Greatness. Chess splits the point as 1.0 for a win, 0.5 for a draw, and 0.0 for a loss. Other modalities of competition may split the point with more or less refinement than chess, but the app will work just fine no matter how refined the point is split. The G-formula is the following:

G = speed of P-Zero gain*resistance coefficient*stamina coefficient*66.7 (34)

The ratings before the tourney begins can be ELO, EDO, Chessmetrics, or even other possible ratings or estimates. The formula doesn't really use these ratings anyway. Rather, it uses the gaps in the given ratings to arrive at the G-score. Thus, this formula, as much as possible, is immune from the effects of rating inflation/deflation. Because of this, this formula gives us a tool to compare the greatness of players from different era's and even different forms of competition.

We shall discuss the components of the formula in turn. First, we have the speed of the P-Zero gain. The P-Zero scores are computed using my algorithm called the PRA. The PRA is an algorithm I have discovered, and it stands for Performance Rating Algorithm. It measures ratings like the Elo system measures ratings, but it does it so much better. In time, the PRA will be the new standard upon which to give ratings.

The P-Zero score is obtained by running through the games played of the tournament with each player's initial rating set to zero. The name P-Zero is one that I coined. It is short for PRA from Zero. Then, the PRA updates these Zero ratings from this strange place of initial player equality as it processes the game data.

At the end of the game data, the P-Zero scores tell us a great deal of information. Players with a positive P-Zero score did better than average in comparison to the field of players. Those players with a negative P-Zero score did worse than average in comparison to the field of players. The player with the highest P-Zero score performed the best, and the one with the lowest P-Zero score performed the worst. We can take the P-Zero scores at the end of the tourney, and use them as a data set representing the performances for the players for this tournament. Now that we have a data set for the players, we can do statistics on it, and the P-Zero scores are at the highest level of data possible for statistical use—the ratio level since there is a natural zero starting point. Ratings themselves are only at the interval level of measurement for statistical use. Thus, the P-Zero scores are really nice things to have.

Nearly 100 percent of the time, no P-Zero scores are duplicated for a tournament. Furthermore, I have never seen an inversion from the P-Zero scores, which is a higher P-Zero score for one player over another in spite of having fewer points garnered in a tournament. All the tie-breaking systems I have seen so far exhibit inversions at times, but not the P-Zero scores—this is true only when each player plays all the games. If there are players withdrawing from a tournament partway through or a lot of forfeited games, then the P-Zero scores are likely to exhibit inversions from what the points

dictate; however, this just shows that the P-Zero scores are better than points. One major example of this is Fischer in Sousse 1967 where he withdrew at about the halfway point. His points were low because he had some forfeits and many missing games due to his withdrawal, but his P-Zero score was second to only Larsens. Fischer's G-score in this tourney was higher than Larsens. The P-Zero score led to Fischer having the second highest performance rating in the tourney, and Fischer was doing phenomenally. In light of this, the tie-breaking order I recommend is points, performance rating, and then G-score. With this protocol, Fischer had too few points to warrant any prize at all, which is as it should have been—he walked out. But, from a peformance perspective, his performance rating and G-score measured him as excelling just behind the winner Larsen, and above the winner Larsen, respectively. So, when the games played are not equal from the players, then inversions can occur, but as this example shows, the performance ratings and G-scores are doing their job correctly, as are the points. So, suffice it to say that the P-Zero scores are better than points. This makes the P-Zero scores ideal as a tie-breaking system for players tied in points.

In *The Greatness of Chess Archive*, a first rank work that searches the great tournaments and matches of chess's past and present, it finds and ranks the greatest performances of all time. The Archive uses the Greatness Formula to render its verdicts.

I rank the players in a tournament precisely by their P-Zero scores. In this way, I show the history of chess as it should have been in The Archive. There are many disagreements between actual history and how the Archive renders it. In real life, the protocol needs to be that forfeited games count as points toward a player accumulating points in a tournament. But, such forfeited games do not count as a game for rating purposes. The rating system should be separated from this. Punishing the player who forfeited the game by getting his rating lowered a little bit is wrong because it does not accurately reflect that he would have lost that game had he actually played it. So, the performance ratings and G-scores are based on games actually played. Points garnered by forfeits can help a player win a tournament but should not have any effect on raising his rating. The integrity of the ratings is a high calling, indeed, and this is one of the things that comprises it (albeit just a little bit because not that many games are forfeited).

I had to give a small bit of background. So, to complete the first item in the G-formula, we need to assess the speed of the P-Zero gain for each player. This is simply done by taking the player's P-Zero score and dividing it by the number of games that that player played. In science, we oftentimes take times and distances to convert to speed. Speed is what we can compare to assess performance. So, using the speed of the P-Zero gain is a great science based way to assess differences in performances between the players. The P-Zero scores are based solely on the played games of the tournament or match, so prior ratings have no effect on their computation. Further, the outcomes of the played games are not used anywhere else in the formula, so this is the only place where that information is used.

Next, we will discuss the resistance coefficient. The resistance coefficient is the average of the individual resistances encountered by the player. The resistance coefficient makes use of the previous ratings (or the ratings that the players have going into the tournament). This is the only place in the formula where infomation about the historical rating of the player is used. Also, the player's own historical rating is irrelevant; we want to assess the resistance that the player encountered—which means we utilize the ratings of the opponents that the player faced to ascertain

the resistance encountered. This resistance is standardized in three ways: firstly, by the second highest rating in the world. Secondly, it is standardized to the constant 2800.0. And thirdly, it is standardized by the calibration of resistance to stamina function. We shall discuss each of these in turn.

Let us say that Rating2 is the second highest rating in the world. Why do we use the second highest? Well, it is an arbitrary thing, really. But, the rationale is that sometimes, the world number one establishes a big gap to his or her nearest competitor. So, by picking the second highest, other world elite players won't get a smaller than deserved resistance that they contribute. So, the first standardization of resistance is to establish a ratio of the rating of a player to the rating of the second highest player in the world. Let us say we have players A, B, C, and D in a tournament. Let us say they play a round robin tournament. So, just looking at player A, he will face players B, C, and D although maybe not in that order. We would observe that the resistances encountered for player A (at this first level of standardization) is RatingB/Rating2, RatingC/Rating2, and RatingD/Rating2. But, as ratings undergo inflation or deflation, these ratios will fluctuate or slide around. So, to give them immunity from this, we have a second level of standardization.

I need a "line in the sand." I need this line in the sand to prevent G-scores in the future from sliding around like the ratings and performance ratings do nowadays. So, I have a formula:

slide = 2800.0 – Rating2. (35)

The slide variable measures the difference that Rating2 is from 2800.0. So, we can now upgrade the resistances offered by players B, C, and D to player A that we showed a moment ago. They are now (RatingB + slide)/(Rating2 + slide), (RatingC + slide)/(Rating2 + slide), (RatingD + slide)/(Rating2 + slide). But, observe that each of these denominators is equal to 2800.0 by rearranging the slide equation. So, we may say that the second layer of standardization produces the following three resistances: (RatingB + slide)/2800.0, (RatingC + slide)/2800.0, (RatingD + slide)/2800.0. So far so good, but there is one more problem!!

We want to average the resistance encountered to arrive at a single number—the resistance coefficient. Nearly 100% of the time, it will be between 0.0 and 1.0, but on rare occasions, it could go over 1.0 by a tiny amount if a player plays the number one rated player in the world repeatedly as in a match. Playing the number two player in the world repeadly will result in a resistance coefficient experienced as exactly 1.0. Other than these two occasions, it will just get tantalizingly close to 1.0 in elite level events. So, what is the problem? The problem is that the resistance has not been calibrated to the stamina. We haven't discussed the stamina coefficient yet. So, let's do that now.

The stamina coefficient is very easy. It is 1.0 + games played/20.0. So, for one game played, the stamina is 1.05. For two games played, it is 1.10. Thus, for each new game played, the stamina goes up by 0.05 points. Note that the stamina coefficient makes use of the fact that a game was played, not its outcome. Whether the player wins or loses (or draws) a game does not deter from the fact that the player struggled and invested energy into the game, contributing to the late tournament fatigue seen by participants. So, we want to institute some kind of reward for simply having played another game that just increments by the same amount no matter what the outcome.

But, this increment at 0.05 seems like such a small amount. Why bother to calibrate that? Well, the answer to that is easy to see from the following possibility. Suppose an elite level player plays a simultaneous exhibition against a field of players each rated exactly 1400. Her stamina coefficient is quietly ratcheting up without bound. Her G-score will mount and mount and mount without end. If she plays, say, 200 players, her G-score will far exceed the greatest G-score I have registered historically in The Archive. This is unacceptable. So, we need one more standardization—to calibrate the resistance encountered to the stamina coefficient.

I have generated some fictitious simuls to get a sense of what the G-scores will be for simuls of various sizes and rating strengths of the opposition (all set to the same value). I am using the calibration function descibed shortly. Here is the table:

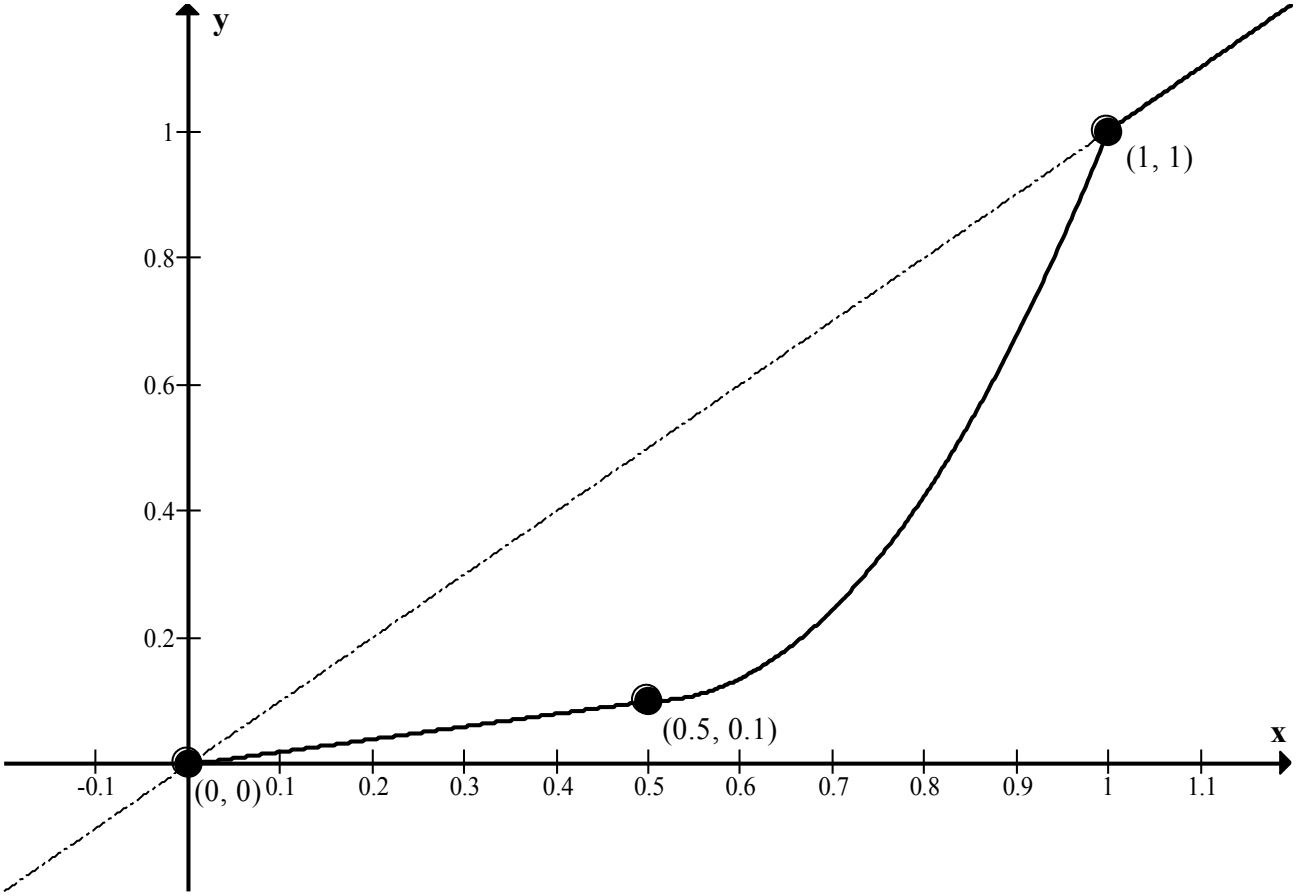
Rating	220 games	500 games	1000 games
1400	172.94	226.98	300.89
1500	180.88	237.40	314.71
1600	204.71	268.67	356.16
1700	244.41	320.78	425.24
1800	300.00	393.74	521.95
2000	458.83	602.18	798.28
2200	681.18	894.01	1185.14

I have concluded that when the resistance coefficient is 0.5 at the second level of standardization, the calibration function f(r) should be 0.1. It may be that over time I could discover that this doesn't work very well. If that is the case, I may need to have a separate G-score formula for simuls if the need arises, but I think tweaking the calibration formula to conform to future observations will be all the adjustment that will be needed. For now, I will use my formula as it is in its current calibration. So, to be mathematical about it, the calibration function is a piece-wise defined function (picture on the next page):

$$f(r) = \begin{cases} 0.2r, 0.0 < r < 0.5 \\ 3.6(r - 0.5)^2 + 0.1, 0.5 < r \le 1.0 \\ r, r > 1.0 \end{cases} \quad (36)$$

The resistance offered by an opponent is (Ropp + slide)/2800.0. Each opponent a player faces contributes to the resistance coefficient. We simply sum these resistances faced and divide by the number of games played to get an overall resistance coefficient (the average resistance faced). Opponents far from player 2 in rating will give a poor resistance individually (i.e. less than 1.0 and sliding down towards zero depending on just how low the rating is). If too many of these players are in the tourney, it will wreck the resistance coefficient and bring down the G-score of the winner of the tourney. A strong tourney will give a resistance coefficient near one. More players far from the rating strength of player 2 will bring down this coefficient. Thus, now that we have the Greatness Formula, top players should be upset at the tournament organizers for bringing comparatively low rated players into the tourney because having to play them means a lower resistance coefficient,

which translates to a lower G-score (provided that the great player wins all the games—not winning all of the games nullifies this argument).



Resistance Coefficient Calibrated to the Stamina Coefficient

Note, however, I have designed a new kind of tournament called the Block Transition Tournament that makes it possible that quite a number of comparatively weaker players can be present in a tourney and still not be able to face the eventual winner of the tourney in a game. I designed this new kind of tournament specifically so that a great player can play the strongest opponents only during the tourney, which provides the greatest possible resistance to the great player. This is done so that if the great player "rises to the ocassion," then he or she will get a great G-score. But, I will talk more about that another time.

The 66.7 is there as a magnifier. It's effect is purely psychological. If no constant was provided, then we would get G-scores below 5. With the 66.7 there, these scores can reach into the 200's and 300's. I have not seen a G-score exceed 400 yet even with about 150 years of data among the rather small number of tournaments I have processed from chess history. Theoretically, there is no upper bound for a G-score. So, it appears that the formula has found that 400 appears to be the peak of human potential for chess. At any rate, the magnitude difference seems more significant with the 66.7 in there, so I keep it.

The task is for a great player to get as big of a plus score as possible. Longer tourneys give rise to the possibility of greater G-scores, but it is called a stamina factor for a reason. The great player

cannot afford a lapse due to fatigue or the G-score will suffer. He cannot avoid facing comparatively lower rated players (in round robin tournaments), so when he does face one, he must defeat him (or her, of course). Occasional draws are unavoidable, but losses really hurt the G-score.

The computations involved are many to process a single tournament, but computers don't mind doing tons of them. Thus, this formula is best done in a computer application. CIS courses could use this paper as an assignment even. I think it is a fine introductory assignment.

I have processed a few tourneys. Here are my findings for G-scores for them (just the 300's—the full list is on Google Drive). Siegbert Tarrasch is the champ so far for tournaments. Wilhelm Steinitz is the champ so far for matches.

1) Steinitz (Steinitz_vs._Blackburne_1876) got G = 397.44
2) Tarrasch (Vienna_1898) got G = 390.21
3) Pillsbury (Vienna_1898) got G = 373.57
4) Fischer (Candidates_1971_Larsen) got G = 369.38
5) Zukertort (London_1883) got G = 368.43
6) Fischer (US_Open_1964) got G = 362.25
7) Lasker (London_1899) got G = 353.87
8) Fischer (Candidates_1971_Taimanov) got G = 350.33
9) Lasker (New_York_1893) got G = 348.19
10) Alekhine (San_Remo_1930) got G = 345.27
11) Beliavsky (Alicante_1978) got G = 333.45
12) Tarrasch (Tarrasch_Waldbrodt_1894) got G = 331.36
13) Korchnoi (Asztalos_Memorial_1965) got G = 328.31
14) Morphy (Morphy_Schulten_1857) got G = 325.69
15) Morphy (1stAmericanChessCongress_1857) got G = 324.25
16) Blackburne (Berlin_1881) got G = 321.56
17) Marshall (Cambridge_Springs_1904) got G = 320.17
18) Capablanca (New_York_Rice_1913) got G = 317.68
19) Kotov (Saltsjobaden_Interzonal_1952) got G = 313.79
20) Fischer (Buenos_Aires_1970) got G = 313.63
21) Karpov (Linares_1994) got G = 308.89
22) Fischer (Palma_de_Mallorca_1970) got G = 307.66
23) Nimzowitsch (Dresden_1926) got G = 307.11
24) Kasparov (Batumi_2001) got G = 305.55
25) Caruana (Sinquefield_Cup_2014) got G = 302.89
26) Kasparov (Tilburg_1989) got G = 302.54
27) Capablanca (Capablanca_vs._Kostic_1919) got G = 302.02
28) Lasker_Em (New_York_1924) got G = 300.11

Next, we shall discuss the Tournament Strength computation. The tournament strength is equal to a weighted average of the resistance brought by each player to the tournament. It is weighted by the number of games each player has played. Of course, if they have all played an equal number of games, then the weighted average becomes just a simple average. But, if there were any forfeits or withdrawals (or if a strange rule was in place to cause an unequal number of games), then the weighted average will account for this discrepency. In symbols using i = 1, 2, ..., n where n is the number of players in the tournament, we have the following (on the next page):

T.S. = 100.0*summation[resistance$_i$*games$_i$]/summation[games$_i$] (37)

Using this formula (where resistance is computed as it is above with the three standardizations), we arrive at the following list of tournament strengths in the archive. I of course had read that AVRO 1938 was one in the discussion of being the strongest tournament ever. I have also read about the 2009 Tal Memorial. Also, some other tournaments in the past have come up as candidates for being the strongest ever. One that I never read about was the Las Palmas 1996 won by Kasparov. Look at the roster of that tournament. All would become world champions with Kasparov as the sitting world champion at the time of the tournament. This formula for computing the tournament strength had this tournament just behind AVRO 1938 and stronger than the Tal Memorial of 2009. So, I looked at the players in this tourney. Of course, the Las Palmas 1996 had an incredible line up with no weak players. So, the computation made me look at this tourney because I didn't believe that a Las Palmas 1996 tournament could be stronger than a Tal Memorial 2009 because I hadn't read about that one. But, upon further review of the situation, the computation result brought it to my attention what an incredible tournament the Las Palmas 1996 tournament was, and Kasparov won it!! So, it gives me pride that my equations do work. So, here is the list of the top tournaments in the archive as of this writing, but just the ones at 90.0 and above.

1) World_Championship_1972: 102.2321
2) Steinitz_vs._Blackburne_1876: 101.4643
3) Kasparov_Deep_Blue_1997: 101.1607
4) Capablanca_vs._Lasker_1921: 100.9464
5) Morphy_Lowenthall_1858: 100.0714
6) Morphy_Anderssen_1858: 99.9060
7) World_Championship_2013: 99.5762
8) Candidates_1971_Larsen: 99.3421
9) Kasparov_Deep_Blue_1996: 98.5858
10) Sinquefield_Cup_2014: 98.4786
11) Candidates_1971_Petrosian: 98.4036
12) AVRO_1938: 97.6703
13) Las_Palmas_1996: 97.5214
14) Botvinnik_vs._Tal_1961: 96.7480
15) Pearl_Springs_2009: 96.5359
16) Candidates_1971_Taimanov: 96.5054
17) Candidates_2013: 96.3398
18) Tarrasch_Waldbrodt_1894: 95.9798
19) St._Petersburg_1895: 95.8885
20) Tal_Memorial_2009: 95.3126
21) M-Tel_Masters_2009: 95.1834
22) Norway_Supreme_Masters_2013: 93.7658
23) Linares_1993: 93.4691
24) FIDE_Grand_Prix_2012: 93.3489
25) Zug_2013: 93.2896
26) Thessaloniki_2013: 92.9658
27) Linares_1994: 92.7961
28) Frankfurt_Chess_1997: 92.7409

29) Lasker_vs._Janowsky_1910: 92.6465
30) Alekhine_Memorial_2013: 92.0589
31) London_Classic_2012: 91.8626
32) Tata_Steel_2011: 91.3134
33) Kamsky_Short_Candidates_1994: 91.1832
34) Baku_2008: 91.1682
35) Steinitz_Martinez_1883: 91.0520
36) Capablanca_vs._Kostic_1919: 90.7590
37) Elista_2008: 90.7504
38) Nalchik_2009: 90.6507
39) Sochi_2008: 90.1728
40) Tata_Steel_2013: 90.1536

Observe that the matches and tournaments here are quite jumbled up from a historical perspective. Capablanca vs. Kostic took place in 1919, and it is sandwiched in between a match involving Steinitz and Martinez in 1883 and a qualifying tournament in 2008. But, the commonality between these three is that there is less than 1% total difference as measured on the scale ranking these events in terms of their strength. By using the three standardizations as described in this essay, the efforts of the past can be compared with the efforts of today. No preference is given to any time period. We get to see an honest comparison of tournaments past and present with this tournament strength score for tournaments and matches.

Here was a special G-score project I came across. As a matter of American dignity, I had to address this. Garry Kasparov sent an email about Nakamura's 2011 Tata Steel performance. He said that it was the greatest performance by an American since Pillsbury in Hastings 1895. He said one had to go back that far to find a performance on a par with what Nakamura had done. He named a few American performances but immediately dismissed them. He gave conflicting reasons why they were to be dismissed. He said some other things, but this is enough to get what he was driving at.

To go back all the way to Pillsbury is to dismiss quite a number of whole generations of American players, especially Bobby Fischer. He took special pains to single him out, which was really unfair. Using my G-score, we can get a proper reckoning of American performance since Pillsbury. I will copy and paste the relevant performances in the list from worst to best. Let us say a G-score within 10 points of Nakamura's Tata Steel 2011 performance is on par, but of course, if it is higher (if any such could be even found Kasparov seems to be implying), then we will include it as well.

137) Dake (Syracuse_1934) got G = 164.70
138) Fine (Syracuse_1934) got G = 164.63
129) Kamsky (Las_Palmas_1994) got G = 169.79
126) Nakamura (Tata_Steel_2011) got G = 170.55
119) Evans (Havana_1952) got G = 177.45
118) Fischer (Candidates_1971_Petrosian) got G = 180.46
117) Pillsbury (London_1899) got G = 180.56
114) Kashdan (Syracuse_1934) got G = 187.59
105) Kostic (New_York_1918) got G = 198.24
103) Sierawan (Wijk_aan_Zee_Hoogovens_1980) got G = 200.51

102) Kamsky (Kamsky_Short_Candidates_1994) got G = 200.56
101) Browne (Wijk_aan_Zee_Hoogovens_1980) got G = 202.23
91) Kashdan (Frankfurt_1930) got G = 213.86
90) Steinitz (Vienna_1898) got G = 215.37
89) Byrne_R (Leningrad_Interzonal_1973) got G = 216.64
88) Pillsbury (Monte_Carlo_1903) got G = 217.96
83) Marshall (New_York_1915) got G = 226.29
80) Lombardy (World_Junior_Championship_1957) got G = 234.89
77) Fischer (Netanya_1968) got G = 240.58
71) Fischer (Sousse_Interzonal_1967) got G = 251.06
69) Pillsbury (Hastings_1895) got G = 253.16
54) Reshevsky (Syracuse_1934) got G = 270.26
48) Fine (New_York_1948) got G = 276.21
43) Reshevsky (Havana_1952) got G = 286.14
41) Fischer (Stockholm_Interzonal_1962) got G = 286.41
38) Hodges (Delmar-Hodges_Match_(Skaneateles)_1892) got G = 286.95
30) Fischer (Mar_del_Plata_1960) got G = 295.29
22) Fischer (Palma_de_Mallorca_1970) got G = 307.66
20) Fischer (Buenos_Aires_1970) got G = 313.63
17) Marshall (Cambridge_Springs_1904) got G = 320.17
8) Fischer (Candidates_1971_Taimanov) got G = 350.33
6) Fischer (US_Open_1964) got G = 362.25
4) Fischer (Candidates_1971_Larsen) got G = 369.38
3) Pillsbury (Vienna_1898) got G = 373.57

A number of these performances are matches, but even so, there is an abundance here of tournaments in the time period Kasparov is talking about with amazing performances by Americans. Yes, some of these are not wins by Americans, but they are "performances" greater than Nakamura's 2011 Tata Steel performance. Yes, Nakamura won the tournament, but his relatively low G-score shows that he did not dominate like many of the performances shown here. Please, Garry, don't insult whole generations of a society to try to make an invalid point.

Another world champion, Bobby Fischer, said once that Reshevsky was the strongest player in the world in 1948 and would have easily defeated Botvinnik in a match for the title. Reshevsky never got the chance due to collusion against Reshevsky by the Soviets. It is a bit of a black eye for chess. However, viewing the list above, Fine and Reshevsky both have 250+ scores, and Fine's is in 1948. But, Reshevsky has one in 1934 and one in 1952 showing he was world class for many years. It's tough to say who would have won. They both were great. I need to process more of chess history to make more definitive statements if anyone else would measure up to these two. I have read that Soviet grandmasters were under orders to throw some games to Botvinnik to ensure his victory in 1948 at the Hague, so a G-score computation for this tournament is suspect, of course.

As a side note, I just found out about Chess960. I do respect Bobby Fischer greatly, but game invention is just not his forte. Tines and Barbs would be something that achieves what he was looking for with a better game. I have also recently found out about Laska—a game invented by Dr. Emmanuel Lasker. I was left profoundly flat by this game. It can be confusing to play because of the

rule that forces certain events on the board. Also, it's complexity level is far too low. I believe computers could play this game perfectly. So, please consider playing Tines and Barbs as it was invented by a real game inventor. It seems playing at the highest levels and inventing at the highest levels are different skill sets. I only have a few case studies to present now, and I will also provide a couple of other essays at the end to complete my work. Thank you for reading my work. I do believe it is the solution to the rating problem as I hope to have shown. If I am successful in this endeavor, many organizations will benefit from using it. Next, I will present some case studies I wrote followed by a couple of essays. Then, I will have a small section on what I would like next for this system.

Case Studies

I have placed the case studies in the archive. See, for example, Havana 1952 and Cambridge Springs 1904. I have provided notes to many of the tournaments in the archive, which also serve as beyond just analysis and give some background or other such things to flesh out the material. Next, I will present two essays I wrote a little while back, but they make some important points and observations about this and other rating systems.

Bobby Fischer Strikes Again

Early in Bobby Fischer's career, he dominated the other players. This was when he was a 12 year old boy. He took so many points from the much higher rated players. The rating system they were using at the time was designed by Kenneth Harkness, and it was called the Harkness system. This was in the 1950's. The Harkness system also gave ratings that were too high according to experts. But, Bobby Fischer's dilemma posed too much of a problem in and of itself. The U.S.C.F. brought aboard Arpad Elo to establish a new rating system to fix this. So, Elo designed his new system. More detail about the downfall of the Harkness rating system can be found at

http://www.avlerchess.com/chess-computer/Chess_Ratings_185258.html.

That system has been in use for over 60 years now. Did Elo get it right? Well, he did a nice job. It was a nice enough job to give it the staying power that it has had. But, his system has problems, too. The main one that is currently plaguing it is called rating inflation. How can we have a number of players today with ratings over 2800 and even Bobby Fischer never exceeded that. The implied argument is that all of these 2800+ players would defeat Bobby over the board. No one who has seriously studied chess history would believe that. For the record, I have solved the rating inflation/deflation problem, but that is another story for another essay (book, actually). But, getting back to this current essay, another seemingly unnoticed problem with the Elo system manifested when Bobby Fischer defeated Boris Spassky in 1972.

In October 1972, we had the ratings of Bobby Fischer and Boris Spassky at 2785 and 2660, respectively. Bobby had established a 125 point gap between himself and the number two player in the world (the champ, Spassky). They played for the championship. The Elo results after that match are as follows:

1) Bobby Fischer: Elo Before W.C.: 2785 Elo After W.C.: 2768
2) Boris Spassky: Elo Before W.C.: 2660 Elo After W.C.: 2677

The simple question to ask is how can Bobby Fischer's rating go down when he won the match handily (12 ½ to 8 ½)? This is a four point margin of victory. We are not talking about a close match here. This is a dominating performance by Fischer. By any objective standard, his rating should increase. But, the Elo formula has him taking a 17 point loss. Spassky, on the other hand, gained 17 points. How can Spassky lose the match by that wide of a margin and still get a rating increase? Well, he shouldn't. It exposes a flaw in the Elo rating system. So, by Bobby Fischer establishing a wide gap between himself and everyone else, he has busted the Elo formula. He

struck first as a boy by busting the Harkness rating system. Then, at the end of his career, he struck again as the new world chess champion by busting the Elo rating system. He has shown that the Elo rating system is flawed.

It took a big gap between his rating and the rating of his opponent (and a dominant performance over that opponent) to show how the Elo system was flawed. Could anyone else ever show this flaw, too? We need someone else to show the flaw or else the Elo system will continue to be used since Bobby Fischer effectively quit playing after he won the title. I do not know if Karpov or Kasparov has lost rating points in this fashion because I have not yet finished my analysis of chess history. I will get to them in due time. But, I have discovered that Magnus Carlsen has lost rating points for winning the Candidates Tournament. So, his gap of 62 points over Kramnik was also enough to show the flaw in the Elo system.

Magnus Carlsen won the Candidates Tournament with a score of 8.5 points. Yes, Kramnik also had 8.5 points, and yes, Kramnik should have won based on the raw performance scores using the PRA (my algorithm). But, Magnus scored 8.5 points also, and that performance was high enough above the field. The tie-break system used by Fide was to give the victory to the player with the most wins. I have no power, so what I say matters not to them. But, the PRA and the raw performance scores will be used for Tines and Barbs players. It is there that justice will be served.

1) Magnus Carlsen: Before Tourney: 2872 After Tourney: 2866.25
2) Vladimir Kramnik Before Tourney: 2810 After Tourney: 2818.58

We see that Carlsen lost about six rating points. No, it isn't quite the 17 that Fischer lost, but he still lost points for winning the Candidates Tournament! How is that possible? It shows what Bobby Fischer showed a number of years ago that the Elo formula is flawed. So, is there a formula that could have handled the situation that Bobby Fischer revealed back in 1972 and that Magnus Carlsen is showing us now? The answer is yes, of course there is such a formula. It is called the PRA, and it stands for the performance rating algorithm. I am working on a book to describe this entire algorithm. The name of the book will be titled Beyond Elo: The Performance Rating Algorithm by Tony Berard. Until then, I can just show results of it. I have processed the 1972 championship match with the PRA. I got the following results:

1) Bobby Fischer: Elo Before W.C.: 2785 PRA After W.C.: 2794.52
2) Boris Spassky: Elo Before W.C.: 2660 PRA After W.C.: 2650.48

So, Bobby Fischer gained about 9.5 points of rating with the PRA for his dominating performance over Boris Spassky. Boris, on the other hand, lost about 9.5 points of rating. This is fair. The PRA is not based on statistical principles. No, rather it is a calibrated rating system to a known population of players. Exactly how I managed to pull this off will all be explained in my upcoming book. Until I get it written, I am just reporting results about tournaments and matches. Regarding the Candidates Tournament that just completed, what would Magnus Carlsen's new rating be using the PRA? Here are the results using Elo and the PRA in a nice table form:

Final standings:
1. GM Carlsen Magnus NOR 2872 – 8.5

2. GM Kramnik Vladimir RUS 2810 – 8.5
3. GM Svidler Peter RUS 2747 – 8
4. GM Aronian Levon ARM 2809 – 8
5. GM Gelfand Boris ISR 2740 – 6.5
6. GM Grischuk Alexander RUS 2764 – 6.5
7. GM Ivanchuk Vassily UKR 2757 – 6
8. GM Radjabov Teimour AZE 2793 – 4

	Start	Elo After	PRA After
1)	2872	2866.25	2879.22
2)	2810	2818.58	2821.94
3)	2747	2765.70	2758.35
4)	2809	2812.13	2816.34
5)	2740	2746.35	2738.62
6)	2764	2764.40	2760.99
7)	2757	2755.54	2750.41
8)	2793	2763.06	2766.13

So, with the PRA, Carlsen gained just over 7 points. Kramnik gained about 12 points with the PRA. No one would quarrel that Kramnik gained only 9 points with the Elo, but we scratch our heads when we see that Carlsen lost rating points with the Elo system when he won the tournament. So, I can confidently report that the PRA can monitor the might of a player such as Bobby Fischer or Magnus Carlsen. They will get the proper credit they deserve for winning tournaments and matches no matter how much of a gap over their peers they might have. There is a wave of expectation going around facebook about Magnus Carlsen. It has his picture on top of a prominent 3000. The post asks the user when did they think that Magnus will reach 3000: by the end of this year or by the end of next year? Well, if the Elo formula won't let a gap get too large, then he cannot get to 3000 unless he has a group of players that he plays against at maybe 2900. Rating inflation is bad, but it isn't that bad. So, unless we scrap the Elo rating system in favor of the PRA, then Magnus will never reach 3000.

Why Kramnik Should Get this Opportunity

I say this with no malice whatsoever towards Magnus Carlsen, but Vladimir Kramnik is the one who should get the opportunity to challenge Viswanathon Anand for the world title. The issue at stake is the tie-breaking system.

The powers that be in chess have decided that the number of wins is the first criterion for what breaks a tie for two or more players that have the same score. This is what is enabling Magnus Carlsen to get a shot at the title. He has five wins to Kramnik's four wins. But, the flip side of this is that Magnus Carlsen has two losses to Kramnik's one loss. So, by them having an equal number of points, a greater number of wins must also entail a greater number of losses. The issue is not debated or considered at all. It is simply more wins = shot at title.

There have been many considerations in the past about how to handle the tie break to figure out who to give a title to or how to distribute the prize money. They are all bad, really. It doesn't have to do with the quantity of options because many have been generated over the years and decades (and

centuries for that matter). The problem lies in the inequality generated by White's first move. It creates an imbalance in the game right from the start. At the highest levels in chess, White wins 22%, Black wins 12%, and 66% of the games are draws. At the Candidates tourney, we saw that nearly with White winning 15 games, Black winning 11 games. Also, 31 games were drawn. This is 26.32 percent White wins, and 19.30 percent Black wins with 54.39 percent draws.

We have had measures such as a rapid tie-break game where White gets a reduction in time, but if Black can win or draw, then that is counted as a Black win. Players still hope for the Black pieces in this scenario because of the huge draw statistic in chess. The disparity in the outcomes of chess do not permit a nice solution in this way.

We have had a measures such as the Neustadtl score and the Buchholz score, but these measures have their critics. The links below are the Wiki's for these scores.

http://en.wikipedia.org/wiki/Sonneborn-Berger_system
http://en.wikipedia.org/wiki/Buchholz_score

So, is there a better way? Well, of course, the answer is yes. I devised a rating algorithm called the PRA. It simply stands for performance rating algorithm. Given a set of pretournament ratings and the games and outcomes of those games, the PRA will calculate the new ratings of the players. Yes, we have had many rating systems before (Elo, Glicko, etc.). What makes me think mine is any better than these? Well, my rating system is not based on statistical principles. Rather, it is a calibrated system to a known population of ratings. It works on populations with different shapes to their distributions. So, it has been tested, and it converges to the known population rather quickly. It has two parts to it: The Basic System, and The Boosting System.

The Basic System is very accurate, but it is slow. It requires a lot of games to be played before it brings the population in line. So, that is why I invented The Boosting System—to speed up the process. But, this essay is not about the PRA per se. It is about a better tie-breaking system. The Basic System can be used to ascertain performance because of its great accuracy. But, we need to set all the ratings to an equal value before processing the games. The ideal value is zero to start everyone at. Then, a positive score would indicate a better than average performance for that tourney, and a negative score would indicate a worse than average performance for that tourney. I have processed the Candidates Tournament 2013 using the Basic System of the PRA with all ratings set to zero at the outset. It produced the following output:

1) 12.98
2) 13.23
3) 9.01
4) 8.59
5) -4.24
6) -4.39
7) -8.53
8) -26.65

Carlsen was player 1 [Note: I have improved the Tournament App since I originally wrote this paper, but it's output formatting still has room for improvement. This is a goal of mine after the Breakthrough Prize contest—to improve many parts of this app.]. He got 12.98 of raw gain in rating. Kramnik (player 2) got 13.23 of raw gain. So, the PRA in its great accuracy is saying that Kramnik outperformed Carlsen. If chess used the PRA, it would be a good decision on its part. Of course, it won't. It will continue to use the Elo system. But, if it did, then Kramnik would be the one going to face off against Anand, not Carlsen. The before and after Elo ratings and the after PRA

ratings for the Candidates Tournament are the following:

	Start Elo	After	PRA After
1)	2872	2866.25	2879.22
2)	2810	2818.58	2821.94
3)	2747	2765.70	2758.35
4)	2809	2812.13	2816.34
5)	2740	2746.35	2738.62
6)	2764	2764.40	2760.99
7)	2757	2755.54	2750.41
8)	2793	2763.06	2766.13

Observe that Elo says that Carlsen should lose rating points even though he won the tournament!! My PRA says that he should gain about 7 points. Kramnik gained a little over 8 points, according to Elo. But, my PRA says he should have gained nearly 12 points. Some other differences between the PRA and Elo exist as shown on the chart above. Notably, player 6 (Grischuk) has Elo giving him a slight increase in rating, but the PRA has him losing about 3 points. My PRA wouldn't punish player 8 (Radjabov) so much, though it is still a pretty steep hit in his rating.

I read an interview that Boris Gelfand had after the tournament. He said that rating doesn't matter. Is he kidding? Of course, rating matters. That is the thing that lets us determine that he is a world class player. He can say that, I guess, because he is elite. He will still be elite because he will keep his rating (a six point gain, according to Elo, but a one point loss, according to PRA). It is rating that gets invitations to the strongest tournaments. Players with lower ratings have to play in lesser tournaments in the hopes of doing well to increase their ratings. I could not disagree with Gelfand more strongly. I really don't think he agrees with himself, either. He may just be saying that to get a fan reaction.

Well, I hope to have convinced readers of this essay that Kramnik should be the one going to the World Championship, not Carlsen. Also, I hope you want to learn more about my work. If so, please join my group Measuring Chess Greatness at http://www.facebook.com/groups/426121960802920.

At some point, there will be a population of Tines and Barbs players. Tines and Barbs is a board game I invented. The tie-break system for tournaments of Tines and Barbs will involve the use of the PRA as I have described in this essay. If this were a Tines and Barbs tournament, Kramnik would be the one going to face off against Vishy.

What is Next?

Well, that is a big question. I will answer it in two ways. The first way is that I need to fix whatever is corrupting my simulation. Then, I will upgrade it to produce more metrics about the PRA. I also want to study other rating systems to be able to code them into the simulation. I want to see how they do with the population of 15000. At some point, I would like a bigger computer so that I can run the simulation on say 10 million players.

I want to study two variable statistics to see if I can apply anything from there into my app to make it even better. It's great already, obviously. But, if it can be improved further, I would like to make that happen. I might even take a course or two in graduate statistics to broaden my horizons.

I hope to have shown in this paper that the PRA and Tournament App are incredibly valuable contributions to organizations needing rating systems and tie-breaking methods that are fair and free of politics. I developed my entire system first through the simulation and then by analyzing the great tournaments of the past with a simple desire to know the truth. I have wanted my system to be free of complaints such as rating inflation, Elo Hell [This is when a good player gets off to a bad start, and Elo gives an initial low rating. It takes an incredibly long time to climb out of this pit, and players that experience this have called it Elo Hell. I read conflicting information about this, though. Some say the players don't climb out simply because they are bad players. Nevertheless, I don't want that sentiment to be attached to my rating system.], and unfair tie-breaking systems. I have achieved this. But, I still haven't definitely answered the question of who was the best player ever. I am closing in on the answer, I think. But, I need the help of a computer scientist to work with me to build something like a Chessmetrics site that taps into a massive database of games. My website, should it be completed, won't have things like missing initial ratings or unusually large performance ratings based on performances against rated players only ignoring the rest of the unrated players (who many times are actually grandmasters in ability or least international masters). A true rendering of chess history would be very nice to accomplish. If I should win the Breakthrough Prize, I will endeavor to complete this massive project. I would be able to hire at least one person to help.

I would like to get my board game off of the ground and establish a population of players for it using my rating system. I would also like my rating system to be used everywhere Elo is currently used. Knocking Elo off of the perch as the world's leading rating system would be really something to be proud of. So, these are some of the things that would be up and coming if I should win. A benefit to the chess world would be quick recognition of grandmasters, like how I identified Sevian and Troff as grandmasters far in advance of them attaining the title through the official means. People that have looked at Sevian's rise in rating from when I had measured him as a grandmaster are impressed that my system did that. In the future, with my rating system in place, Sevian would be awarded the grandmaster title at the conclusion of the tournament. This can become the norm if I can achieve what I envision as the future for this rating system.

About the Author

My name is Tony Berard, and I invented the board game Tines and Barbs. I also wrote a couple of books. One of them is *The Flaws, Death, and Successor to Chess: Tines and Barbs*, ISBN 978-1456562526 available on Amazon.com. This book is my inventor's statement about my board game Tines and Barbs and why it is a superior game to chess. It includes all there is to know about the game at least at this stage in the game's history.

The other book I wrote is a math book called *Points, Lines, and Conic Sections: A Sequel to College Algebra* ISBN 978-1456377199, also available on Amazon.com. This book has quite a

number of new equations related to the conic sections. It requires that the student has had college algebra, and it builds on the student's knowledge to include my new equations for the conics. I didn't think that the material I developed for my discoveries in the conics was enough for a college course, so I integrated my material with material from a subject in mathematics called Triangle Geometry. It is now a full length college level course in mathematics. One day I hope that this book becomes a new college course for math majors as well as other scientific disciplines needing more robust mathematical knowledge about the conics than usual.

After these two projects, I began to develop the PRA and Tournament App. This paper presents the seminal findings, but I still have a long way to go before I can begin to consider this project done.

Along the way in life, he acquired a bachelors degree in mathematics from Lawrence Technological University in Southfield MI. He currently lives in Waterford MI.