A Comprehensive Guide To Chess Ratings

Prof. Mark E. Glickman
Department of Mathematics
Boston University
111 Cummington St.
Boston, MA 02215
mg@math.bu.edu

Introduction

The introduction of chess rating systems has probably done more to influence the increase in popularity of tournament chess than any other single factor. Even though the theory of the current rating system, often called the "Elo" system after its creator, was developed as far back as the late 1950s, the International Chess Federation (FIDE) only adopted the rating system in 1970. Since that time, the system has been adopted with various modifications by many national chess federations. Today, it is impossible to imagine tournament chess without a rating system.

Chess rating systems have several practical uses. For pairing purposes in tournaments, a tournament director should have some idea which players are considered the most likely candidates to win the tournament so the director can effectively avoid pairing them against each other during the earlier rounds of the tournament. Ratings are also used for tournament sectioning and prize eligibility; a section in a tournament may only allow players of a specified rating range to compete for section prizes. Ratings can also be used as a qualifying system for elite tournaments or events; invitation to compete in the U.S. closed championships and to compete on the U.S. olympiad team are based in part on players' U.S. Chess Federation (USCF) ratings. The current "title" systems used by some chess federations base their title qualifications on the overall strength of tournament participants as measured by their ratings. But probably the most useful service of the rating system is that it allows competitors at all levels to monitor their progress as they become better chess players.

The first chess rating system to produce numerical ratings was the Ingo system developed by Anton Hoesslinger in the Federal Republic of Germany in 1948, and named after his home town, Ingolstadt. Over the next ten years, various forms of this system were used by different national chess administrations, including versions developed in the mid-1950s for the USCF by Ken Harkness and for the British Chess Federation by Richard Clarke. These systems combined the frequency of winning with the level of opposition. While these Ingo-based systems were popular in the 1950's because the ratings they produced were consistent with subjective rankings of chess players, they had little basis in statistical theory. In fact, in the Harkness system, a player could lose every game in a tournament and still gain rating points. This and other flaws in the Harkness system yielded way in the U.S. for the adoption of the Elo system in 1960.[1]

The Elo rating system assigns to every player a numerical rating based on performances in competitive chess. A rating is a number normally between 0 and 3000 that changes over time depending strictly on the outcomes of tournament games. When two players compete, the rating system predicts that the one with the higher rating is expected to win more often than the lower rated player. The more marked the difference in ratings, the greater the likelihood that the higher rated player will win.

While other competitive sports organizations (U.S. Table Tennis Association, for example) have adopted the Elo system as a method to rate their players, non-probabilistic methods remain in use for measuring achievement. In the American Contract Bridge League (ACBL) bridge rating system, "master points" are awarded for strong performances. Points are awarded relative to the playing strength of the competitors in an event. For example, the number of master points awarded to a bridge partnership in a national championship compared to that in a novice tournament could be as high as 750 to 1.[2] One of the key differences between the Elo system and the current ACBL system is that the Elo system permits a rating to increase or decrease depending on a player's results, while the bridge system only allows a rating to increase, and never decrease. A bridge rating is therefore not only a function of one's ability, but also a function of the frequency in which a player competes. Because of this characteristic, bridge players' abilities cannot be directly compared via their ratings. Ratings derived under the Elo system, however, are designed, in principle, to permit such a comparison.

Another system that has gained acceptance is one of several used for rating professional tennis players. For example, the ATP tennis ranking system awards "computer points" based mainly on the type of tournament (e.g., "Grand Slams," "Championship Series," etc.), total prize money in

[1] The first published description of the system appeared in "New USCF Rating System," Chess Life, June, 1961, 160–161.

[2] This figure was provided by Alan Oakes, Director of Member Services at the ACBL.

the tournament, and the highest round a player attained before being eliminated (or if the player won the tournament). Players are ranked by the sum of the computer points corresponding to their best 14 results from the previous 52 weeks, or the sum of all the computer points if competing in fewer than 14 tournaments. This system, like the ACBL bridge rating system, does not have probabilistic underpinnings, but does seem to produce rankings that roughly correspond to popular belief. Unlike the bridge rating system, repeated poor performances can decrease one's ATP rating. The ATP system also adds the element of time into the system, which is lacking in both the Elo and ACBL systems. The Elo and ACBL systems use a player's most recent rating as the current rating even if the player has not competed in a long time, whereas in the ATP system a player can lose rating points through a lack of competition. This feature may be more appropriate for tennis than for chess or bridge where one's ability may be more clearly linked to one's frequency of competition. A curious feature of the ATP system is that tennis ratings can change abruptly. For example, if a player has won a major event, and during the following year has mostly mediocre results, then at the year anniversary of winning the major event the player's rating can be expected to drop precipitously. So while the ATP system does incorporate a time component to ratings, it does not guarantee smooth changes in rankings.

This paper describes the basic principles of the Elo rating system, and how these principles are manifested in currently implemented rating systems. Throughout the paper, the USCF rating system serves as the focus of attention, though much of the discussion extends to other implementations of the Elo rating system.

Statistical context

The problem of rating chess players falls into the topic of "paired comparison" modeling in the field of statistics. Paired comparison data results from any outcome that indicates a degree of preference of one object over another. Clearly, chess outcomes fall into this framework because a chess game is the result of two players being "compared" to determine who is the "preferred" player (or whether "no preference" is made, in the case of a draw). Other examples of paired comparison data occur in other sports whose results are wins and losses, e.g., football, basketball, and hockey. The outcomes of these games can also be viewed as indicating a degree of preference through score differences; games where one team defeats another by a large margin conveys a greater degree of preference over a game where the final score difference is close. Also, topics in experimental psychology such as choice behavior and sensory testing involve paired comparison data. For example, the "Pepsi challenge" is a test to determine whether an individual prefers Coca-cola to Pepsi-cola. Iowa State University statistician Herbert David provided a good overview on statistical modeling and analysis of paired comparison data in a 1988 monograph.[3]

While Elo's name is by far the most associated with the development of the current chess rating system, the statistical development underlying the system had been established well before his work in the late 1950's, and certainly prior to his well-known 1978 monograph.[4] The first work to give serious attention to modeling chess ability was E. Zermelo in 1929.[5] In this paper, Zermelo addresses the problem of estimating the strengths of chess players in an uncompleted round robin tournament. Statistician I. J. Good in 1955 developed a system that amounted to the same model as Zermelo's, but was obtained through a different set of assumptions.[6] Both of their models are connected to the Bradley-Terry model for paired comparison data, which was first described in detail in a paper by statisticians Ralph Bradley and M. Terry in a 1952 paper.[7] Among popular paired comparison models, the Bradley-Terry model has the strongest connection to the currently implemented Elo rating system.

One way to understand the Bradley-Terry model, or most other models for paired comparison data as they relate to chess, is to suppose that every player brings a box containing many numbered slips of paper when sitting down to a chess game. Each number represents the player's potential strength during the game. This collection of values will be called a player's "strength distribution." Instead of actually playing a chess game, each player reaches into the box and pulls out a single piece of paper at random, and the one containing the higher number wins. In effect, this model for chess performance says that each player has the ability to play at a range of different strengths, but displays only one of these levels of ability during the game. Naturally, this procedure favors the person who carries a box that contains generally higher numbers, but of course this does not necessarily imply an automatic victory. This is analogous to chess where a better player usually wins, but not always.

The Bradley-Terry model can be derived by making a particular assumption about the frequency distribution of values in player's box. If every player's strength distribution (i.e., distribution of values in the player's box) follows what is called an "extreme value distribution," then the Bradley-

[3] *The Method of Paired Comparisons*, (Oxford University Press, 1988).

[4] Arpad E. Elo, *The Rating of Chessplayers, past and present*, (ARCO, 1978).

[5] "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Math. Zeit.*, 29, 436–460.

[6] "On the Marking of Chess Players," *Mathematical Gazette*, 39, 292-296.

[7] "The Rank Analysis of Incomplete Block Designs. 1. The Method of Paired Comparisons," *Biometrika*, 39, 324–45.

Terry model results. The shape of the extreme value distribution is shown in Figure 1. The height of the curve at a particular strength value describes the relative frequency a player will randomly select that value. For example, because the curve is roughly twice as high at a strength of 1500 relative to 1300, a player with the extreme value distribution in Figure 1 is twice as likely to perform at a strength of 1500 compared to a strength of 1300. Under the Bradley-Terry model, every player's distribution of strength follows an extreme value distribution having the same shape, but centered at a different value depending on the player's overall ability. Note that the curve trails off more slowly to the right, so that the assumption of an extreme value distribution implies that a player is more likely to randomly select a high number from his or her box than a low number. Thus the Bradley-Terry model postulates that a player will play with an ability that fluctuates from game to game, but rarely will the ability be substantially lower than one's average display of ability.

Because we are primarily interested in the likelihood one player will defeat another, it is just as important to consider the distribution of the differences between randomly selected values from each player's box. The proportion of the time that the difference is greater than 0 tells us the probability one player will defeat another. The Bradley-Terry model assumes that if we consider all possible combinations of values from one player's strength distribution with all possible combinations of values from an opponent's strength distribution, the differences between the two numbers over all these combinations follow a "logistic" distribution. This distribution is shown in Figure 2. Under the Bradley-Terry model, the probability that the first player will outperform the other is the fraction of the area under the logistic curve that is to the right of 0. This is exactly equivalent to the first player having drawn a higher value from his or her strength distribution.

Even though the currently implemented system can be derived by assuming that a player's strength distribution is an extreme value distribution, Elo's development of the chess rating system assumes that the a player's strength distribution is a normal distribution (bell curve). Figure 3 shows the curve for the normal distribution. The paired comparison model that is derived from the normal distribution is commonly known in the statistics literature as the Thurstone-Mosteller model, based on work by mathematician L. Thurstone in the late 1920's,[8] and statistician Fred Mosteller in the early 1950's.[9] Psychometricians William Batchelder and Neil Bershad in 1979, using the Thurstone-Mosteller model, extend Elo's model by formally modeling the probability of individual game outcomes.[10] One interesting feature of using the normal distribution to model a

[8]L. Thurstone (1927) "A law of comparative judgment," *Psychological Review*, 34, 273–286.

[9]F. Mosteller (1951) "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, 16, 3–9.

[10]Batchelder, W. and Bershad, N.(1979) "The statistical analysis of a Thurstonian model for rating chess players,"

player's strength distribution is that if we consider all combinations of values from one player's strength distribution with all possible values from an opponent's strength distribution, the differences have the same shape, though the differences are more spread out. The distribution of differences appears in Figure 4.

It appears as though there is very little distinction between the shape of the logistic distribution in Figure 2 and the normal distribution in Figure 4. Figure 5 shows both curves superimposed, with the logistic distribution drawn as a solid line. In fact, statistics professor Hal Stern in a 1992 article[11] has shown that when analyzing paired comparison data, it makes virtually no difference whether one assumes the logistic distribution or the normal distribution for differences in players' strengths. So, empirically, the choice between the Bradley-Terry model and the Thurstone-Mosteller model is a moot issue. Mathematically, however, the Bradley-Terry model tends to be more tractable to work with. This is the most likely reason that most organizations administering a probabilistic rating system (e.g., FIDE, USCF) use the Bradley-Terry model, which uses the logistic distribution assumption, rather than the Thurstone-Mosteller model, which uses the normal distribution assumption.

More recent development of models for rating chess performance have appeared in the statistical literature. Statistics professor Harry Joe in a 1990 paper examined the world's best chess players of all time by a model that splits players' careers into "peak" periods and "off-peak" periods.[12] This analysis was performed on a data set compiled by Raymond Keene and Nathan Divinsky, a prototype of which appeared in their 1989 monograph "Warriors of the Mind."[13] Statistician Robert Henery in a 1992 paper analyzed this same data set, and proposed using the length of a game to predict the outcome of chess games.[14] In a more developmental approach, Harry Joe wrote an article in 1991 that derived axiomatically a general framework for a rating system, and showed how the Elo system is a special case.[15] A recent article by Batchelder, Bershad and R. Simpson[16] uses a "reward system" approach, similar to Joe's, to updating players' ratings.

Paired comparison theory has most typically been devoted to problems modeling judges' preferences among a set of objects. While the game of chess, and most other games involving two

Journal of Mathematical Psychology, 19, 39–60.

[11] Hal Stern (1992) "Are all linear paired comparison models empirically equivalent," *Mathematical Social Sciences*, 23, 103–117.

[12] "Extended use of paired comparison models, with application to chess rankings," *Applied Statistics*, 39, 85–93.

[13] *Warriors of the mind: A quest for the supreme genius of the chess board*, (Hardinge Simpole, 1989).

[14] "An extension to the Thurstone-Mosteller model for chess," *The Statistician*, 41, 559–567.

[15] "Rating systems based on paired comparison models," *Statistics and Probability Letters*, 11, 343–347.

[16] Batchelder, W., Bershad, N., and Simpson, R.(1992), "Dynamic paired-comparison scaling," Journal of Mathematical Psychology, 36, 185–212.

competitors, can be viewed as a paired comparison insofar as a player is "preferred" when he or she wins a game, what makes the problem of rating chess players different from the usual paired comparison setting is that players' abilities can and do change over time. This is a non-trivial aspect to the problem. My Ph.D. thesis in 1993 develops an approach for solving this problem. In my work, I described a general probabilistic mechanism by which players' abilities change over time. As an application of the models developed, I analyzed the results from the World Cup chess tournaments of 1988–9 to determine ratings of the participants in the events. The approach I have taken to modeling change in abilities over time was independently formulated by German statisticians Ludwig Fahrmeir and Gerhard Tutz,[17] though my approach to data analysis is slightly different. Though Elo's development of a rating system does not go into the same level of mathematical detail as in my dissertation, perhaps Elo's most significant contribution was the introduction of a simple algorithm to adjust players' ratings from tournament game outcomes. Elo's framework is quite appealing: players have ratings prior to a tournament which, in theory, predict their performances, game outcomes are observed, and players' ratings are adjusted to account for the differences between the observed results and the pre-event expectations. This process is then repeated for the next event. While much of Elo's development can be criticized for its lack of reliance on established statistical principles, he successfully implemented a system that appears to track players' performances with reasonable adequacy.

Ideas underlying the Elo rating system

Rating parameters and rating estimates

When statisticians analyze data with the hope of explaining or understanding the mechanism by which the data are generated, they make a very clear distinction between "parameters" and "estimates." To understand the difference, consider the following situation. Suppose one is interested in finding out the proportion of tournament chess players in the U.S. who believe that Fischer could defeat Kasparov in a 24-game match. This proportion, which is a characteristic of the population of U.S. tournament chess players, is an example of a "parameter". Its exact value can only be known by obtaining the opinions of every tournament chess player in the U.S. To find the precise value of this parameter would be absurd. One would need to ask the opinions of tens of thousands of players in order to learn the answer. Even if the means were available to ask everyone, one

[17] Fahrmeir, L., Tutz, G.(1994), "Dynamic stochastic models for time-dependent ordered paired comparison systems," Journal of the American Statistical Association, 89, 1438–1449.

is probably not interested in knowing the parameter value with such precision. Instead, a more convenient approach would involve gathering a small sample of players, and guessing the parameter value based on information from the sample. To accomplish this, one might randomly select 200 players from all over the country and ask their opinions on a potential Fischer-Kasparov match, and compute from this sample the proportion who believe Fischer would win. This value computed from the sample is an example of an "estimate" of the parameter. The proportion who believe Fischer would win calculated from the sample of 200 players is expected be close to the proportion calculated from the entire population of tournament players (if such a task could possibly be carried out), so a great deal of work has been saved by calculating an approximate answer. On the down side, the value calculated from the sample would likely be different if one were to obtain a different sample of 200 players. So, for example, it may be possible to randomly choose a sample of 200 players of which 42% believe Fischer would win, and then randomly select another sample of 200 players of which 35% believe Fischer would win. This reveals the main drawback of relying on estimates; they are subject to variability. The tradeoff is clear – the more accuracy we want in estimating a parameter, the greater the expense (usually in the form of acquiring a larger sample). The usual role of a statistician in this type of situation is not only to estimate the parameter value from a sample, but also to understand how much the estimate can be expected to vary from sample to sample, and to identify a reasonable sample size so that estimates are not likely to vary much from sample to sample.

The distinction between estimates and parameters is rarely, if ever, made in the context of chess ratings. For a true appreciation of the rating system, this distinction is important to understand. Returning to the analogy of players drawing numbered slips of paper to determine the outcome of a game, one might be especially interested in the average value of these numbers for a particular player. The Bradley-Terry model (and therefore the systems used by the USCF and FIDE) assumes that the only difference across players in the distribution of the numbered slips of paper is their center or average (because the spread of values around the center is assumed identical). An examination of the left plot in Figure 2 makes this point clear. The two superimposed curves represent the frequency of values from two players' strength distributions. The only difference between these two curves is that the curve drawn as a solid line is shifted to the right relative to the curve drawn as a dotted line. This suggests that we only need to keep track of the center (average value) of each distribution, because that is the only feature of the two distributions that is different. Once we know the average value of a player's strength distribution, we should be able to describe the entire distribution of values. It is this average value or average strength, a parameter that is a feature of a player's strength distribution, that we want to learn about in a chess rating system.

Unlike the previous example where it is merely inconvenient to find out the exact proportion of players who think Fischer will defeat Kasparov, it is actually impossible to learn the exact value of the center of a player's strength distribution. The reason can best be understood by analogy to the previous example. To discover the proportion of chess players that believe Fischer will defeat Kasparov, one needs to identify the population of interest, and then specify the computation that leads to the parameter value. This is a straightforward procedure; one could conceivably list every member of the tournament chess playing population, ask each person his or her opinion, and then produce the value of the parameter by dividing the number of players that believe Fischer would win by the total number in the population. In the chess rating situation, the "population" would be considered all possible displays of playing strength (i.e., all numbered slips of paper from a box). If one could possibly have knowledge of such information, then we could somehow compute the average across an infinite number of values to obtain the average value of the player's strength distribution. Clearly, it is impossible to observe even a single value, much less a collection of values, from a player's strength distribution. Instead, only games outcomes can be observed, so an estimate of a player's strength parameter must somehow be inferred from a sample of game outcomes. This estimate of a player's average strength is what we know as a chess rating.

A computed chess rating is really an estimate of the player's rating parameter, that is, the player's average strength.[18] To understand the connection between a reported chess rating and a rating parameter, consider the following situation. Suppose a player has a strength distribution with an average value of 1654, even though this could not possibly be known. When this player registers for the tournament, the tournament director finds that his reported rating from the most recent rating list is 1693. In this particular instance, the player's estimated rating of 1693 is higher than his true, though unknown, rating parameter of 1654. This suggests that this player can be expected to perform worse than his published rating would lead one to believe. This example points out that because published ratings are merely estimates of rating parameters, they are subject to variability and imprecision. A player's published rating would likely be a different value had the player competed against different opponents in his or her last tournament. We may also conclude that, just as in estimating the proportion of all players who think Fischer could defeat Kasparov, the more often a player competes the more precise we are likely to estimate the player's average strength.

Ironically, however, the fundamental mathematical assumption of the USCF and FIDE rating systems involves a statement about the rating parameters, and not about the ratings that are

[18]The terms "rating parameter" and "average strength" are synonymous and will be used interchangeably throughout the discussion.

printed in rating lists. In a game played between players with true average strengths of R_A and R_B, the expected score of the game for player A is assumed to be

$$E = \frac{10^{R_A/400}}{10^{R_A/400} + 10^{R_B/400}}, \tag{1}$$

where the score of a game is 1 if player A wins, $\frac{1}{2}$ if the game is a draw, and 0 if player A loses. The expected score of a game has an interpretation as a long-run average. If players A and B were to play repeatedly, assuming their abilities do not change, then the average of the scores corresponding to their game outcomes will be close to E. Suppose, for example, that the rating parameter for player A is 1500 and the rating parameter for player B is 1700. Then the above formula states that the expected score of the game for A is about 0.24. This implies that player A will win at most 24% of his games against player B in the long run, and probably less than 24% because many of these games will be draws. The paradox, of course, is that this formula applies only to rating parameters, which we never can know exactly, and not to estimated ratings, which are computed based on observed data. Suppose, in the previous example, that the published rating estimate for player A is 1547 and for player B is 1661. If we blindly applied the expected score formula pretending that these values were the true parameter values, we would falsely conclude that the expected score of the game for player A is 0.34, a value which is substantially larger than the value computed using the exact parameter values of 1500 and 1700.

One might be tempted to think that the differences between estimated ratings and rating parameters would average out when computing the expected score; some players will have an estimated rating that is greater than their rating parameters, and other players will have lower estimated ratings. Interestingly, an analysis of the outcomes of over 8300 USCF-rated tournament games demonstrates that the expected score function computed on estimated ratings does not describe the data. The game results were taken from several tournaments between 1991 and 1993, including the 1992 U.S. Open, the 1993 National Open, the 1991 and 1992 Illinois Open events, and the 1993 Los Angeles Open. Figure 6 shows the results of the analysis. The games were grouped according to the players' differences in their published USCF ratings at the time of the events. The figure shows the average score for the higher rated player for various rating differences, along with a 95% margin of error.[19] The dotted line in the figure corresponds to the expected score according to the formula in (1). If estimated ratings were interchangeable with rating parameters, then the dotted line would intersect the segments on the figure. In most cases, the expected score overestimates the observed average score for particular rating differences. This suggests that either

[19]The 95% margin of error for the average score provides a plausible range of values for the average score if new data were gathered. In particular, the range of values in the figure would occur in approximately 95% of new samples. Shorter segments indicate, to some extent, larger samples.

the formula assumed in (1) is not correct, or the rating estimates are not good approximations to the rating parameters.

At first, this consistent overestimation of the expected score formula may seem surprising. In truth, if a rating parameter is estimated with error from player to player, we *should* anticipate that the expected score formula overestimates the observed outcomes. This is actually a statistical property of the expected score formula. To understand this, suppose, for example, that the rating estimates for every player in our sample were determined randomly so that a player's reported rating would have no connection to a player's true average strength. In that case, if we were to reperform the analysis that led to Figure 6, we should expect all the average scores for each rating grouping to be centered close to a horizontal line at 50%, as the randomly determined rating provides no information about the players' abilities. At the other extreme, if rating estimates were so precise that they were exactly equal to rating parameters, then we would observe the expected score curve intersecting all the segments. What we actually do observe is something in between these two extremes – the segments are centered somewhere between 50% and the expected score curve. This simply implies that estimated ratings are not meaningless (or else the segments would be very close to a horizontal line at 50%), but they are not exact either (or the segments would intersect the expected score curve). Fortunately, the figure indicates that the segments are closer to the expected score curve than they are to 50%, especially at the higher rating differences.

Another way to understand this overestimation is to consider what happens when a player with a true average strength of 1900 plays against an opponent with a reported rating of 1700. Suppose that the reported rating of 1700 is imprecise, so that approximately one half the time the player plays at an average strength of 1600 and the other half of the time plays at an average strength of 1800. If we calculate the expected score using the opponent's reported rating of 1700, we obtain a value of 0.76. In truth, we can expect a score of 0.64 when the opponent plays at a rating of 1800, and expect a score of 0.85 when the opponent plays at a rating of 1600. So, on average, the first player can expect to score

$$(0.64 + 0.85)/2 = 0.745$$

against the opponent. This value is less 0.76, which is the result computed on the reported rating of 1700. Thus the expected score computed on the reported rating is higher than what should actually happen. The mathematical fact illustrated here is that the expected score computed on the average of opponents' ratings is systematically greater than the average of individual expected scores when the opponents' ratings are generally lower. This is likely to be the main explanation for the behavior in Figure 6.

Updating ratings

Because it is impossible to learn a person's rating parameter exactly, the only hope is to be able to accurately estimate the parameter. Suppose a chess player has just finished playing in a tournament. What approach should be taken to estimate the player's average strength? One approach would be to estimate the rating parameter based on game outcomes only from the tournament. An estimate of a player's rating parameter from a single tournament is often called a *performance rating*. This idea seems reasonable, but it ignores potentially useful information from past tournaments. Another approach involves examining the entire history of this player's tournament performances and estimating his or her rating parameter as if all of these games were played in one large tournament. While this makes use of a player's historical information, this has the drawback of treating a recently played game and a game played years ago as equally indicative of current average strength. The most reasonable approach seems to be a compromise between these two extremes. The best estimate of current ability should make use of all tournament games ever played, but should give substantially greater emphasis to more recent games. In effect, this is how the Elo updating formula works.

The rating update formula involves adjusting a player's estimated rating as new data is observed. The adjustments are made incrementally so that rather than recomputing an estimated rating from a player's entire tournament history, a pre-tournament rating is used as a summary of his or her history prior to the current tournament. This allows for a simple recursive description of the rating procedure; a player's post-tournament rating is a weighted average of an estimated performance rating with an estimated pre-tournament rating. Because calculating performance ratings accurately involves a computation that can be too demanding to perform on a regular basis, an approximation is used. The formula for adjusting a pre-tournament rating is given by

$$r_{post} = r_{pre} + K(S - S_{exp}), \tag{2}$$

where r_{post} is a player's updated post-tournament estimated rating, r_{pre} is a player's estimated pre-tournament rating, S is the player's total score in the tournament, S_{exp} is the expected total score estimated from the player's pre-tournament rating and the player's opponents' pre-tournament ratings, and K is an attenuation factor that determines the weight that should be given to a player's performance relative to his or her pre-tournament rating. The term S_{exp} can be calculated by summing the expected scores, E, for each game using formula (1). Of course, this is only an approximation to S_{exp} because in using formula (1) the estimated ratings are being substituted for the rating parameters.

The above formula has some nice interpretations. First, the term $(S - S_{exp})$ can be thought of as a discrepancy between what was expected and what was observed. If this term is positive, then the player performed better than expected because the attained score, S, is greater than the total expected score, S_{exp}. Therefore this player is likely to be stronger than the pre-tournament rating predicts, so the player's rating is increased by the discrepancy magnified by the value K. Similarly, if the term $(S - S_{exp})$ is negative, then the player must have performed worse than expected, and therefore this player's rating will decrease by the discrepancy magnified by the value K. The larger the discrepancy, $(S - S_{exp})$, in magnitude, the less "valid" the pre-tournament rating must have been, and the greater the change required to properly adjust the rating. For example, if a player were expected to score 3 points out of a 5-round tournament given the opponents' pre-tournament ratings and proceeds to lose every game, then the pre-tournament rating was a poor predictor – it should have been much lower to produce such a lackluster performance. When $(S - S_{exp})$ is zero, then the player's expected score is exactly equal to the attained score. This suggests that the player's pre-tournament rating correctly predicts the actual performance in a tournament, so no adjustment is required. It is worth noting, however, that these calculations assume the opponents' reported pre-tournament ratings are known and are accurate estimates of their average strengths.

The attenuation factor K in formula (2) can best be interpreted as the amount of weight given to the new tournament performance relative to the pre-tournament rating. The larger the value of K, the greater the amount of change allowed in one's rating. It can be shown mathematically that for a 4-round tournament, setting $K = 32$ corresponds approximately to computing a weighted average of a pre-tournament rating and a performance rating with weights equal to 94.7% and 5.3%, respectively.[20] This implies that each time a new tournament is observed, 94.7% of our belief is invested in the old rating, but we let 5.3% of our belief be guided purely by what happens in the tournament. If computing a tournament performance rating, r_{perf}, were a straightforward calculation, then an alternate method for computing a post-tournament rating corresponding to $K = 32$ would be

$$r_{post} = 0.947 r_{pre} + 0.053 r_{perf}.$$

Analogously, when $K = 24$, the weights become 96.2% and 3.8%, respectively, and when $K = 16$ the weights become 97.5% and 2.5%, respectively. These approximations only hold when the discrepancy $(S - S_{exp})$ is not too large.

An analogy can be drawn between the formula in (2) and tracking the position of a moving

[20] The mathematical justification involves an approximate relationship between the quantities $(S - S_{exp})$ and $(r_{perf} - r_{pre})$, where r_{perf} is the rating at which the sum of expected scores equals the attained score. The value that multiplies $(r_{perf} - r_{pre})$ in the formula provides the necessary information to determine the weighting.

target in preparation for firing a missile. Suppose we have a rough idea about the current location of a target, and we aim our missiles accordingly. The laws of physics tell us precisely where the missile is expected to land. The target now moves, and our tracking instrument tells us the approximate location of the target. We can adjust the aim of our missiles to account for this new information. This is analogous to targeting a player's chess ability. A player's pre-tournament rating roughly conveys current playing strength, or the player's "position." The expected score formula summed against his opponents is how the laws of the rating system tell where the "missile will land." An actual total score is observed, and we adjust our "aim" of the player's true "position" by using the formula in (2). The rating system can therefore be viewed as a device that constantly tracks a player's ability as it changes.

Elo's approach to adjusting ratings by equation (2) generally works well when a player's pre-tournament rating is not too different from the player's actual strength. Mathematically, the approximation in (2) as a weighted average between the player's pre-tournament rating and performance rating breaks down when the pre-tournament rating and performance rating are far apart. This could occur if, for example, a player has not competed in a long time. Another instance where it does not make much sense to directly apply the formula in (2) is when a player has never competed in a tournament, so no pre-tournament rating exists.

Provisional ratings

The formula in (2) describes the procedure to estimate a player's rating given his or her estimated pre-tournament rating. This formula would appear to be of little use when a player has no rating prior to entering a tournament. Both the USCF and FIDE have implemented systems to compute initial ratings using a different set of formulas. The resulting estimated ratings are often called "provisional ratings." As the name implies, we do not attach a great amount of confidence in provisional ratings because they are estimates of rating parameters based on a very small sample of game outcomes. A provisional rating in the USCF rating system is an estimated rating that is based on fewer than 20 games. FIDE uses provisional rating formulas to calculate a player's rating during the 6-month period in which the player first competes. Both of these methods involve averaging performance ratings over tournaments for the period a player's rating is considered provisional. With the current implementation of the USCF rating system, this is a problem. Because no time limit is imposed on the duration one's rating is provisional, and because all game results count equally toward one's provisional rating, a game result from a year ago would have the same effect on his or her current estimated rating as a game played in the past week. This could be a problem if newcomers to tournament chess earn a low rating after their first tournament, become discouraged, and then return to tournament chess only after having improved.

An approach that has a strong connection to the rating update formula in (2) can be employed to compute provisional ratings. The idea is simple. Before a player competes in a USCF tournament, he or she is assigned a rating based on, say, age. We'll call this rating a player's *prior* rating, and it is understood that this estimate is subject to a great amount of uncertainty because it is not based on the results of a player's game results. Now this player competes in a tournament, and the formula (2) is applied using the prior rating as r_{pre}, and setting the attenuation factor K to be very large (e.g., 150) to give substantial weight to the performance. For a 4-round tournament, $K = 150$ corresponds approximately to maintaining 38.7% belief to the prior rating and the remaining 61.3% belief to the rating information learned from the tournament game outcomes.

A logical question to ask would be why not simply give 100% belief to a rating computed solely from information from the first tournament. After all, this is the approach both FIDE and the USCF currently use in their computations, and it certainly seems reasonable to base conclusions about a player's ability exclusively from game outcomes. A subtle reason exists for making use of prior information in this context. In statistics terminology, the use of prior information addresses a phenomenon called "regression to the mean," or more generally, "shrinkage."[21]

The idea behind shrinkage can be illustrated by an example. Suppose a group of twenty chess players, all possessing the same average strength, compete in a single round robin tournament, and, for sake of argument, the winner achieves a score of 14 points out of 19. Suppose also that the player with the worst results obtained a score of 4 out of 19. It should not be surprising that one player out of 20 scored as many as 14 points, and that one player out of 20 scored as few as 4 points even though all the players are of the same caliber. Now if these twenty players were to compete in a second single round-robin tournament, it is likely that the results of the winner from the first tournament will not be as impressive as his or her outstanding performance from the first tournament. It could happen, but it is much more likely the player will produce results closer to an average score. Similarly, the player with the worst performance from the first tournament will probably have a performance that is not as poor. In general, it is arguable that players' performances in the second tournament will "shrink" towards the mean score compared to performances in the first tournament. This is not true in every instance; it is just true on average. We can carry this argument directly over to the calculation of performance ratings. When we

[21]A good non-technical introduction to the concept of shrinkage can be found in the Scientific American article "Stein's Paradox in Statistics" by Bradley Efron and Carl Morris, May 1977.

calculate an estimated rating for the player who has won the first tournament, we need to realize that performing a calculation that only uses information from the tournament is likely to produce an *overestimate* of his or her true ability (and analogously an underestimate for a player with a poor performance) because the player has likely over-performed relative to his or her true ability. A way to bring this overestimate back down is to calculate a weighted average of this extreme performance with that which an average player would perform. Naturally, a substantial amount of weight would still be placed on the performance relative to the prior information. This procedure of shrinking values computed solely from the data (e.g., a performance rating) to the prior mean in order to draw conclusions from data is standard in statistical practice, and can directly be applied to the method of rating chess players. As a player continues to compete, repeated use of the updating formula guarantees that the original "prior" rating will have little impact on a player's current rating.

Rating system implementation

USCF and FIDE rating scales

The method Elo laid out for adjusting ratings was adopted by the USCF in 1960, and subsequently adopted by FIDE in 1970. Through the years, various modifications were made to the systems which were catered to the needs of the governing organizations. Originally, the two systems were intended to produce ratings that were meaningful on the same scale. Because the two systems function independently and incorporate slightly different updating algorithms, it is not surprising that a FIDE rating will not correspond exactly in meaning to a USCF rating.

The FIDE scale, which rounds its published ratings to the nearest multiple of 5, only computes ratings as long as they remain higher than 2000. A distribution of the July 1994 FIDE rating list appears in Figure 7. The mean rating for this time period is 2262 which is shown on the figure as a solid vertical line. The proportion of players with FIDE ratings less than 2200 is about 23%.[22] The ratings range from 2005 through 2780. One of the main differences between the FIDE rating algorithm and the true Elo updating algorithm is that Elo's calculation computes the sum of a player's expected outcomes against each opponent, whereas the FIDE algorithm computes the expected outcome against the average rating of the opponents. Mathematically, these two computations do not produce identical results. The FIDE calculation, as Elo mentions,[23]

[22] FIDE only recently allowed all players to acquire ratings less than 2200, so this figure is of some interest.
[23] See section 1.66 of *The Rating of Chessplayers*.

is an approximation to computation that was intended. The calculation carried out by the FIDE algorithm is problematic because if a player competes in an event against opponents with a wide array of abilities, the FIDE calculation may be a poor estimate of a player's total expected score.[24] Another issue concerning the FIDE rating system is that a player only acquires a rating if it is calculated to be over 2000. This suggests that, on average, initial FIDE ratings overestimate players' abilities because players only receive ratings if their initial performances are strong. A player a bit weaker than 2000 strength might have a good performance which would give him or her a FIDE rating, but a player who is stronger than 2000 who has a poor performance would not receive a FIDE rating. Thus the FIDE rating pool has a tendency to inflate over time because the initiated FIDE players tend to decline slightly to their appropriate level while their opponents respectively increase in rating.

The USCF rating system, which assigns ratings to all competitors in USCF-governed tournaments, does not restrict a player to demonstrate strong ability to earn a rating.[25] Thus the range in USCF ratings is much larger than the range for FIDE ratings. Figure 8 shows the distribution of players with established ratings (players with more than 20 rated games) for July 1994. The mean rating for established USCF players in July 1994 is 1490. USCF established ratings ranged from 45 to 2763. About 96% of all USCF established players have ratings less than 2200, as compared to the 23% for FIDE.

A very common misconception about the rating system is that the distribution of ratings across players has some theoretical distribution, such as the normal distribution.[26] No such assumption is made in the Elo system, or in any paired comparison model. The distribution of ratings is merely a function of the strengths of the players that compete. The Elo system only makes an assumption about the distribution of potential strengths an individual might display in a game, that is, the distribution of numbered slips in a player's box. This is an assumption about the range of strengths displayed by a single person, not an assumption about the range of average strengths across players.

An average conversion can be established between the USCF and FIDE rating scales by examining the ratings of players common to both systems. From the July 1994 FIDE and USCF rating lists, a total of 484 players were identified as being common to both. Among these 484 players, only players that had established USCF ratings and had played at least 6 FIDE-rated games in

[24] For example, if a player rated 2005 competes against opponents rated 2600, 2600, 2600, and 2005, then the player can be expected to score about 15%, whereas using the FIDE computation would result in an expected score of about 7%.

[25] The lowest rating a player can earn under the USCF rating system is 0.

[26] For example, a December 1979 Chess Life & Review article, "Ratings – Some Questions Answered" by Gerry Dullea made such a mistake.

the prior six months before the publication of the FIDE rating supplement were included in the analysis. This resulted in a total of 211 players meeting this restriction criteria.

Figure 9 shows a plot of the USCF ratings against the FIDE ratings for the 211 players, with a curve traversing the center of the points. The curve was determined using a statistical technique called "locally weighted scatterplot smoother" that ignored unusual points (e.g., the 2300 FIDE-rated player with the 1800 USCF rating). Apart from some points corresponding to players with unusually low USCF ratings, the pattern of data appears smooth and tightly clustered around the curve, except for FIDE ratings lower than 2200. Figure 10 magnifies the relationship by plotting the FIDE ratings against the USCF–FIDE differences. The curve shows that the FIDE–USCF difference varies according to FIDE rating. For low FIDE ratings, the expected difference between FIDE and USCF ratings is high: the USCF–FIDE rating difference for a FIDE rating of 2050 is about 120; for a FIDE rating of 2100 the difference is about 70. This difference drops down to 30 at a FIDE rating of 2200. The difference climbs again to about 80 for a FIDE rating in the mid-2500's, and then declines once more to a difference of 65 to 70 in the high-2600's. A possible reason that the USCF–FIDE differences are higher for FIDE ratings less than 2200 is that only players with USCF ratings over 2200 play frequently enough (more than 5 games in 6 months) to appear in the analysis. Among the 273 players with USCF ratings that played 5 or fewer FIDE-rated games in the previous 6 months, the USCF ratings tended to be much lower compared to the corresponding players who played more than 5 FIDE-rated games. This may be explained by the earlier argument that newcomers to the FIDE pool of players may be initially overrated.

The PCA rating system

The Professional Chess Association (PCA) has developed a rating system that calculates ratings on the same scale as USCF and FIDE ratings. The pool of players that are rated under the PCA system has large overlap with the FIDE pool, so it can be viewed as a separate algorithm to rate the abilities of the same player population. Ken Thompson of Bell Laboratories was the main force behind the system, with some advice from statistician Axel Scheffner of Germany, economist Andrew Metrick of Harvard University, and me. The PCA system produces ratings for active international competitors. Only the top 500 players in any PCA ratings list are currently published, though all players competing in PCA-rated events possess ratings. The system was originally set up so that the top 150 players in the PCA system were forced to have the same average rating as the top 150 players on the FIDE list.

Every PCA player has either a provisional rating or an established rating. Provisionally rated players are those that have competed in fewer than 25 games against established players. The PCA rating system saves the outcomes of the most recent 100 games in which a player was involved, except that the results against provisionally rated opponents are discarded. A calculation is then performed for each player which estimates the player's rating parameter based on the stored game results (up to 100 games) along with the opponents' pre-event ratings at the time a game was played. The 100 games are weighted "linearly," implying, for example, that a player's 10th most recent game receives 5 times as much weight as the player's 50th most recent game. Games played in the same event receive equal weight. Once these estimates are obtained, the system then calculates a "variance" for an individual player, which is a measure of how erratically a player performs against his or her opponents.[27] The details for the "variance" computation involve calculating the average squared deviation of each game result (1, $\frac{1}{2}$, 0) from its expected game result using the expected score formula, and then transforming this value back to a value interpretable as a rating. This computation of the "variance" addresses the possibility that the box of numbered slips of paper may vary in spread from person to person, an assumption not made in the Elo system, and not assumed in the Bradley-Terry model. However, the PCA algorithm is carried out by first computing rating estimates assuming the Bradley-Terry model (i.e., the "variances" are all the same), and then acting as if each player has possibly different "variances." The result of this procedure are values that are difficult to interpret, except in an ad hoc fashion. A more statistically sound procedure would derive the "variance" measures simultaneously with the rating estimates. Fortunately, the computed "variances" are not used in algorithm to update ratings, so the "variance" computation is not relevant to the predictive ability of PCA ratings.

Fundamentally, the PCA rating algorithm is similar in principle to the Elo algorithm.[28] The outcome of a game follows the Bradley-Terry model, and ratings are updated based on outcomes against opponents along with the opponents' pre-event ratings. The main underlying difference between the two system is in their methods of downweighting past performances. Because the PCA system downweights games linearly, it is difficult to interpret the weights. Consider a player who currently has competed in 100 PCA-rated games. In computing the player's current rating, the outcome of the player's 5th most recent game was given four times as much weight as the player's 20th most recent game. However, after the player has competed in an event consisting of 10 games, the 20th game before the event has now become the 30th game, and the 5th game has now become

[27]The term "variance" has a specific technical meaning in statistical language, and is not used properly by the PCA system. The most obvious disparity in definitions is that a true variance is measured on a scale of squared units, whereas the PCA "variance" is measured on the same units as the rating.

[28]The PCA algorithm does, however, incorporate the advantage due to playing white. This is discussed in a later section.

the 15th game. This implies that the rating calculation weights the more recent game (now the 15th) by only twice as much as the less recent game (now the 30th). It seems counterintuitive to have the weight between games depend on the number of games having been played. The Elo system, by contrast, essentially performs "exponential" weighting which preserves the weighting among events by their respective placement in the order of being rated.[29] This may be an area for possible improvement in the PCA system.

It should be noted that the Elo approach to rating adjustment and the PCA approach share the same basic assumptions, though they are implemented differently. In both systems, previous results are downweighted relative to recent results. The PCA system uses computations that make fewer approximations than the USCF or FIDE systems. This by no means suggests that the USCF or FIDE systems are less accurate. In fact, rating systems that use the Elo updating scheme, such as the FIDE and USCF systems, are following an approach almost universally endorsed by the statistics community. The idea is this: Rather than save all the game results from the past and compute a rating based on all the data each time a tournament is completed, extract only the relevant information from a tournament and use it in combination with the pre-tournament summary information to produce a post-tournament summary. At this point in the procedure, the tournament data may be discarded. This approach recognizes that only certain aspects of the data are relevant for making conclusions about playing strength, so it is not necessary or desirable to save all information and recompute ratings from scratch.

Rating system characteristics

Varying time controls

One of the newer features of the USCF rating system stems from the formal introduction of quick chess. Quick chess now refers to games where the time control for a game is quicker than 30 minutes per person for the entire game. In the late 1980's, a debate emerged whether games played in chess tournaments with fast time controls should be rated under the same rating system that governs ratings for games played under slow time controls, or whether a separate rating scale should be created. Eventually, a second rating system that parallels the original system was constructed to rate these performances separately.

[29]The Elo updating formula is effectively a linear approximation to exponential weighting. This is different from linear weighting, however.

The main argument for using a separate system is that people who perform substantially better at quick chess than at slow chess may be demonstrating a different ability than what is required for winning a slow game. For example, one could argue that a greater number of tactical mistakes are made in quick chess, so players who are comfortable playing double-edged lines may have better performances in quick chess. Because a different ability is being measured, a different rating scale is justifiable. Advocates of separate scales could claim that keeping a single scale for quick and slow chess would contaminate the system in the same way as what could happen if the rating systems for over-the-board and correspondence chess were merged.

Opponents of separate systems for quick and slow chess would probably respond by asking why draw such a solid line at 30 minutes? A player's ability surely is not noticeably different when playing under a time control of 29 minutes for the entire game versus 30 minutes. It is also not obvious that 30 minutes has any special meaning. Why not, for example, draw the line at 15 minutes, or at 45 minutes? These are questions that the advocates for separate systems need to answer before standing on firm ground.

A compromise between these two approaches, suggested to me originally by Roger Cappallo of MIT, involves constructing two rating systems that correspond to time limits of, say, 5 minutes for an entire game and 40 moves in 2.5 hours. When a player competes in a tournament where the time control is in between these two rates of play, *both* ratings are updated. The magnitude of change for each rating depends on the closeness of the actual tournament time control to the time controls of 5 minutes per game and 40 moves in 2.5 hours. Under such a system, a player may approximate his or her rating at various time controls by taking appropriate weighted averages of the two ratings. Of course, this requires a further conjecture about the weights attached to the two ratings, so implementing such a system may be difficult in practice.

Regional variation in ratings

The title of the recent play by John Guare, "Six degrees of separation,"[30] refers to the theory that every two people are connected by at most six other people in the sense that the first person knows A who knows B who knows C, etc., who knows F who knows the second person. The claim, therefore, is that a path can always be traced from person to person that only requires at most six people in between.

[30] *Six degrees of separation*, (Vintage Books, 1990)

In measuring chess ability, this notion of being able to trace paths that connect players has direct relevance. While no claim will be made that any two players have competed via six degrees of separation, the claim can be made that the fewer the degrees of separation between two players, the more accurate the comparison of abilities. For example, most players would probably agree that weekend tournaments attract roughly the same players, so that these local players compete amongst themselves fairly regularly. The ratings for these players are likely to be accurate predictors of how each will fare against the other, assuming one is willing to believe the expected score formula in equation (1). Even in cases where two players have not competed directly against each other, they may each have a number of opponents in common which establishes a connection between them (via one degree of separation). By contrast, when two players live in separate parts of the country where they are not only likely never to have competed, but also have rarely played opponents in common, or even opponents of opponents in common, the accuracy of their ratings as predictors of a game result between the two is put into question.

One of the fundamental problems with using the rating system as a predictor of performance is that it is only accurate on a within-region level. No provisions exist in the rating system to prevent disparities in abilities across different regions of the country for similarly rated players. As an extreme example of how the rating system could provide misleading interpretations, consider two groups of tournament players who only compete among themselves, each of whom have an average rating of 1500. Also suppose the abilities of the players in the first group improve faster than those in the second group. If the players in either group only compete among themselves, then we cannot possibly determine that the players in the second group are better players on average than those in the first group through their ratings. A player rated 1500 in the first group will likely be notably worse than a player rated 1500 in the second group. Some connection is needed between the two groups in order to recognize a difference in abilities.

A situation in which a group only competes among themselves occurs frequently in scholastic chess. At the beginning of their chess careers, scholastic players may only compete against other scholastic players. A community of scholastic players is formed, and very rarely do players venture outside this community to play against adults, and if they do, they rarely return to their scholastic community. The ratings for these scholastic players have an especially poor connection to ratings of adult players because the ratings were first derived from competitions among unrated scholastic players. The ratings for these players, therefore, are poor predictors of performance when they begin competing in adult tournaments.

While most situations are not as extreme as in the preceding examples, they do pose real challenges for a rating system. If communities of players do not compete against each other with any frequency, then the possibility exists that the strength implied by ratings in one community may become different from the strength associated with the same ratings in another community. This leads to the claim by certain cities that they are systematically underrated relative to players in other cities.

The only true remedy to this problem is to ensure that players in different communities compete regularly. This function is served by large state and national tournaments which provide players an opportunity to compete against opponents they would otherwise never encounter. These tournaments can be viewed as big mixing bowls, where the discrepancies among players' ratings relative to their strengths are combined and smoothed out. When they finish the tournament, they bring back to their communities a slight adjustment in their ratings that reflect the overall strengths of their opponents in other communities. A similar adjustment occurs when players move from one residence to another. Such players mix the abilities described by their ratings with the abilities of the players in the new community. The net effect is an averaging of the discrepancies due to regional variation in ratings, although this may not be enough to completely solve the problem.

Time variation in ratings

One of the most natural uses of the rating system is to monitor one's progress over time. Usually, players enter the rating pool with a low rating, and as they gain more tournament experience, their ratings increase slowly and steadily reflecting their improving ability. But is it really the case that an increase in one's rating always connotes improvement?

Relating increases or decreases in one's rating over time to change in ability is very tricky business. Even though one's rating may be changing, it is not clear whether it is changing relative to the entire pool of rated players. As Elo argued, the average rating among rated players has a general tendency to decrease over time. His argument of "rating deflation" involves examining the flux of players into and out of the player population. If no new players enter or leave the pool of rated players, then every gain in rating by one player would (ideally) result in a decrease in rating by another player by an equal amount. Thus, rating points would be conserved, and the average rating of all players would remain constant over time. But, typically, players who enter the rating pool are assigned low provisional ratings, and players who leave the rating pool are experienced players who have above-average ratings. The net effect of this flux of players lowers the overall average rating.

Rating deflation can be defined more specifically as the result of a mechanism that causes players' ratings to decline over time when their abilities, on average, do not decline. Elo's argument for rating deflation can be made tighter. Specifically, the existence of rating deflation requires two features of the rating system. The first is that players' abilities, on average, improve over time. It should not be obvious that this happens because older players may have abilities that are decreasing over time. The second requirement is that the rating system, on average, does not systematically add or subtract points to players' ratings independent of their performances. If these two conditions are met, then there is a tendency for reported ratings to decrease over time even when certain players' average strengths remain constant. These players, in all likelihood, will compete against underrated opponents who are improving, and will on average obtain lower ratings at the expense of the underrated players.

In the mid-1970's, it was becoming apparent that the average rating of USCF players was beginning to decline. Deflation was not only evident from the year to year movement in the average USCF rating, but also from a greater discrepancy in USCF and FIDE ratings. Throughout the past two decades, the updating formulas for the USCF rating system have been modified to combat this rating deflation. One approach was the introduction of bonus points and feedback points in the mid-1970's. When a player performed exceptionally well, his or her rating not only increased according to the usual updating formula, but also increased by the addition of a "bonus" amount. The justification for awarding bonus points was that the player was most likely a rapidly improving player, so the ordinary updating formulas did not track the player's improvement quickly enough. When a player was awarded bonus points for an exceptional performance, the opponents would receive additional points to their ratings called "feedback" points. The rationale for awarding feedback points was that the player's opponents should be rated against a higher pre-tournament rating because the player who was awarded bonus points was notably stronger than his or her pre-tournament rating suggested. To account for this discrepancy, extra rating points were added to the opponents' ratings. By the mid-1980's, these features were eliminated from the rating system, in part because it appeared as though bonus points and feedback points were overcompensating the natural deflationary tendency of ratings by causing the average to increase, and in part because the bonus point and feedback point system had no firm statistical foundation.

In the late-1980's, the concept of a rating floor was established in the USCF system. In its original form, this addition to the rating system prevented a player's rating from decreasing below the 100-point multiple 200 points less than one's highest attained rating.[31] If, for example, a

[31] The highest attained rating for every player was only recorded after the inception of the rating floors.

Rating Status		January 1993	January 1994	Mean rating	Number of
1993	1994	Mean	Mean	increase	Players
Established	Established	1632.6	1641.7	9.1	12233
Provisional	Established	1143.1	1184.4	41.3	1910
Inactive	Established	—	1421.8	—	4393
Provisional	Provisional	1124.7	1138.3	13.6	1772
Inactive	Provisional	—	990.4	—	10777
Established	Inactive	1548.4	—	—	9670
Provisional	Inactive	1086.4	—	—	7933

Table 1: USCF Rating summaries for January 1993 and January 1994

player's highest attained rating was 1871, then the player's rating could not decrease below 1600. More recently, the rating floor has been re-implemented so that now instead of using a 200 point margin, the system uses a 100 point margin. In the example above, under the current system, the player with a highest attained rating of 1871 cannot decrease below 1700. Proponents of the rating floors argue that this will not only combat the natural tendency of rating deflation, but will encourage chess tournament participation because it prevents one's rating from decreasing without limit. Furthermore, the rating floors may discourage players from purposely losing games to artificially lower their ratings which would enable them to compete in lower-rated sections.[32] Nonetheless, the use of the rating floor is at odds with the principle that ratings are measures of performance. Additional rating points are being injected into the system through players who are presumably getting worse rather than those who are getting better, so any inflationary effect of floors is indirect. Furthermore, players at their rating floor may have misplaced incentives, and may therefore adjust their style by purposely playing more recklessly in the hopes of winning against higher rated opponents with less effort. If ratings are to be used as a predictive tool, the rating floor implementation must be considered a flaw in the rating system.

It is interesting to examine changes in the overall rating USCF pool. The USCF publishes annual rating lists which include players who had tournament games rated over the past year. In the January 1993 list, the mean rating among players with established ratings was 1595.4, whereas in the January 1994 list, the corresponding mean was 1542.5. This suggests that the rating pool experienced an average decrease of about 53 points in 1993. Such a simple analysis is misleading, however. Table 1 shows mean ratings broken down into players' statuses in 1993 and 1994. For example, the first line of the table indicates that 12233 players had established ratings in both January 1993 and January 1994. The average rating for these 12233 players in January 1993 was

[32] This is usually called "sandbagging."

1632.6, and this average rating increased to 1641.7 in January 1994. Thus, among players with established ratings in both years, an *increase* occurred in the overall average rating. Table 1 also shows that among players who were provisionally rated in January 1993 and then established in January 1994, the overall average rating increased by 41.3 rating points. Furthermore, players who were provisionally rated in both January 1993 and January 1994 experienced an average rating increase of 13.6 rating points.

How can the overall average rating among established players in January 1993 (1595.4) decrease to the average rating among established players in January 1994 (1542.5) if the average rating among players who were established in both years increased by 9.1 points? The answer lies in the flux of the established rating pool. By the end of 1992, 21903 players had established ratings who were active over the previous year. Slightly more than 44% of these players became inactive in 1993. These players had an average established rating of 1548.4, as shown in Table 1. In contrast, 18536 players had established ratings in January 1994 who were active in 1993. Of these, slightly more than 34% were either inactive or had provisional ratings in January 1993 (corresponding to the second and third rows of Table 1). The average established rating for this group in January 1994 was 1349.9. In addition to maintaining 12233 players from January 1993 to January 1994 who experienced a 9.1 average rating increase, the established rating pool lost a group of players with an average rating of 1548.4, and gained a group of players with an average rating of 1349.9. The net effect of this trade of players into and out of the rating pool resulted in an average rating decrease of 53 points.

The average increase of 9.1 points among players who had established ratings in January 1993 and January 1994 can be shown to be "statistically significant," which implies that the increase is not simply due to random fluctuation in individual ratings. An examination of data from other years leads to the same conclusion.[33] Most likely these established players' ratings increased at the expense of provisionally rated or unrated players, because the updating formula in equation (2) suggests that whenever two established players compete, the gain in one player's rating will result in the other player's loss. The only exception to this occurs when the value of K in the updating formula is different for the two players, but the overall effect of this exception will not make a substantial impact on the overall average rating increase for established players. The other possibility is that, for some of these players, the rating floor has prevented their ratings from decreasing.[34] The magnitude of this effect is difficult to estimate.

[33]A similar analysis was performed on data between 1988 and 1989, and between 1992 and 1993, resulting in the same conclusions.

[34]In January 1994, approximately 8% of all active players with ratings between 1400 and 2200 were at their rating

If the rating system were functioning properly, we should not expect a significant increase in established players' ratings from one year to the next. In particular, the 9.1 average rating increase among this group suggests either that the rating floor is having a sizable effect on the ratings of established players, or that the provisionally rated opponents of these established players are overrated, on average. This second reason can be justified in the following manner. Clearly, provisional ratings are subject to a great amount of uncertainty, so that sometimes one would expect a provisional rating to overestimate a player's ability, and sometimes one would expect it to underestimate. But if the system worked properly, the number of provisionally rated players whose average strengths were overestimated would be at most the number whose average strengths were underestimated. This can be argued realizing that provisionally rated players are ones who are generally improving, so that their ratings are generally underestimates of their true average strength. Assuming this were true, then among all contests involving a provisionally rated player and an established player, the average rating change among established players should be either close to 0 or negative. The intuitive reason is that the rating gains by the established players, who will usually have higher ratings than the provisional opponents, will be relatively small, but will be balanced by the large rating losses when they lose games. So Table 1 may provide evidence that the rating system may not be properly functioning.

Even though adjustments to the rating system have been implemented to counteract rating drift, being concerned about changes in average rating of the tournament chess playing population inherently depends on the goals of the rating system. The rating system by itself only makes assumptions about *differences* in players' ratings, not in the actual value. If 1000 were subtracted off (or added to) everyone's ratings, the rating system would still be just as valid because differences in players' ratings would still remain the same.

Clearly a rating has more interpretive value if it can be understood without directly comparing it to other ratings. When a player talks about being "1800-strength," he or she is doing so with the implicit understanding that a rating of 1800 connotes a specific level of ability. Moreover, a general belief exists that "1800-strength" this year should connote the same ability next year, five years from now, and twenty years from now, and if somehow this does not happen, then something is wrong with the rating system. The fact is that a rating system solely based on game outcomes of players whose abilities may be changing over time is unable to guarantee that a particular rating will connote the same ability over time. This is an observation pointed out by writer and computer

floor. This can be estimated by counting the number of players whose established ratings have 00 as the last two digits and comparing to the number of players with different ending double digits.

consultant John Beasley,[35] who asserts that ratings can only be used to describe relative abilities and not absolute abilities. The abilities of players in the overall population are constantly changing due to factors such as studying, increased understanding of the subtleties of the game, and aging, and this prevents measuring absolute changes in ability from game outcomes. Suppose, for example, two players with 1500 ratings play a 10-game match, each scoring 5 out of 10. This results in post-match ratings of 1500. Now a year goes by, and suppose both players have immensely improved their chess playing ability in the same amount. They compete again, and each scores 5 out of 10 again. Even though both players have improved, we cannot detect this because their ratings will remain unchanged at 1500. Thus, a change in one's rating is only relative to the change in the average ability of the population of players over time.

Recognizing that the rating system may not be able to provide an absolute measure of chess ability, but rather only a measure relative to other rated players, we can identify several possible goals for maintaining characteristics of the overall rating pool. One possible goal is to force the median rating or some percentile of all active players to a prespecified rating by periodically adding a fixed amount to all ratings. Suppose, for example, that a median rating of 1500 is desired. Then 50% of all players will have ratings above 1500 from year to year. This allows a player to compare one's rating with an average rating to determine his progress. A related idea involves specifying a certain small proportion of players to have a rating higher than some threshold value, and periodically add an amount to all ratings to guarantee this. One such rule could be to guarantee that only 1% of all active players have ratings above 2200, and uniformly adjust ratings so that this condition is met. As long as we are consistent in defining what is meant by an active player, then either of these two approaches seems justifiable. Of course, this approach suggests that a player's rating may change from the overall pool adjustment even when he or she is not competing.

Another idea that has been proposed is to align a rating scale to match another rating scale that may be considered more universally acceptable. For example, the USCF has often considered aligning its rating scale with the FIDE scale, and to periodically update USCF ratings so the two scales have the same absolute interpretation. However, neither the FIDE system nor any other system in existence guarantees any particular stability in its rating scale or its rating system. With the recent moves that excluded Gary Kasparov and Nigel Short from the FIDE rating lists, it would appear that the FIDE rating system should not be viewed as the gold standard of rating systems. A further argument against aligning two rating scales, such as the USCF and FIDE scale, is that the link from one scale to the other may be based on a small number of players, so the alignment

[35] *The Mathematics of Games*, Oxford University Press, 1989, pg. 60.

may fluctuate primarily due to the imprecision of the estimated conversion between the two scales. Also, in trying to gain control in the rating system, it is unappealing to impose a condition on a rating system that depends on criteria from another system over which the first system has no control.

Finally, one possible direction of effort is to develop techniques to make ratings connote the same ability over time through external means. Even if the rating system alone cannot recognize improvement in ability, other methods can be incorporated to assess the magnitude of ability changes. One basic idea borrows from "item response theory" in educational testing. The Scholastic Aptitude Test (SAT) taken by a substantial proportion of high school juniors and seniors had been constructed so that current students' performances would have direct comparability to to students' performances of the past. This is accomplished by including a number of test items common to different exams. Thus individual exams are "linked" together by common test questions. Due to these links, paths can be inferred that connect students of the past to students of the present via statistical models. A particular current SAT score would connote the same ability as in the past.[36]

This framework can be applied to rating chess players in several different approaches, though the merit of any of these approaches is certainly arguable. One idea is to make use of chess-playing software. Because the chess-playing ability of a non-learning chess program only improves if the code is revised, a chess program can be viewed as having a fixed ability. To use chess programs for assessing change in ability, the ratings of several chess programs can be accurately estimated by having them compete against each other, as well as having them compete against a wide selection of humans. These ratings can then be used as fixed "anchors" in the rating system. Periodically, these chess programs should be entered into tournaments and the results of competition would determine the magnitude of an overall ratings drift. The drift can then be adjusted by adding or subtracting a fixed amount from everyone's rating. This idea makes the vital assumption that players do not learn how to improve their play against chess software, which is a demonstrably poor assumption as shown in certain test conditions. However, if the chess programs were required to compete infrequently, players would not necessarily have the opportunity to learn how to play against the software. A compelling argument against this approach is that humans play differently against chess programs than they do against other humans. A performance against chess programs may not be representative of what would happen against humans of ratings equal to the chess programs. Also, implementing such a procedure of having programs compete against humans regularly may be impractical and expensive.

[36]As of 1994, the SAT no longer attempts to connect scores to the past in this manner, but instead determines scores that correspond to percentiles of the current population taking the exam.

A variation of this theme is to periodically identify groups of players who seem to demonstrate stable abilities, and use them as anchors in the rating system for a temporary period of time. Essential to this idea is to prevent players from being aware of players used as anchors in the rating system. Players who would be candidates for anchors are those who play regularly, and whose ratings do not fluctuate much. Such players would be used in their capacity for, perhaps, six months at a time, after which the entire rating pool would be adjusted to reflect drift away from these players' ratings. The main criticism here is that it is difficult, if not impossible, to assess *a priori* that a player's ability has reached equilibrium and is not improving.

Finally, a somewhat more scientific approach for ratings connoting the same ability over time involves designing a chess test to measure chess ability, and then fitting a statistical model to predict chess ratings from the chess test. A series of chess questions could be constructed to test ability in all phases of the game. A sample of rated chess players would be administered the test, and formulas could be developed that predicted their ratings with reasonable accuracy merely from the responses to the test questions. This test could then be administered a year later on a different sample of players to check how different the ratings derived from the test results differ from the actual tournament ratings. Based on these differences, an adjustment could be applied to all ratings to restore the constancy of ratings over time. This approach, while making use of an external source to measure chess ability other than game results, has the benefit of identifying the aspects of chess that separate weak chess players from strong ones. On the down side, assessing the accuracy of the test is now a new source of variability, and could increase the difficulty in measuring playing strength. In any case, designing and administering such a test in addition to the statistical analysis of the data could be a costly procedure to carry out correctly, and may not be in the interests of chess organizations.

Improving the rating system

The Elo rating system as it is currently implemented appears to function reasonably well. Even though aspects of the rating algorithm are open for criticism, it is a self-correcting system. If a player's rating fails to represent his or her true average strength, then the rating system will correct the player's rating from the results of tournament competition. Nonetheless, the rating system can be improved in various ways to provide more accurate predictions of performances without having to wait for the system to correct inaccuracies. We examine some areas that are open for improvement.

Advantage due to color

One of the more commonly understood features of chess is that having the white pieces conveys an advantage. Elo estimates that white has a 1.33 times better chance of winning than black.[37] In my Ph.D. thesis, I estimated from results of the World Cup tournaments of 1988–89 that among top masters white has a 1.56 times better chance of winning than black, given players of similar abilities. This corresponds approximately to an 80 rating point advantage for white. With such a large advantage to white, it seems that incorporating color information makes sense.

The advantage for playing white can be framed in terms of randomly selecting numbered slips of paper from each player's box of numbers (strength distribution). When one of the players sits down to play white, the value of 80 is automatically added to every value in the white player's box. This is a very straightforward mechanism to describe how a statistician might model the advantage to playing white.

The rating system can properly account for color by reexpressing the expected game score formula so that color is incorporated. A possible formula for the expected score of a game played between A and B, when A has white, could be given by

$$E = \frac{10^{R_A/400}}{10^{R_A/400} + 10^{(R_B - C)/400}}, \tag{3}$$

where C is the rating advantage conferred to white (C is the number added to every value in player A's box). For example, if two players had the same value of their rating parameters, and C were equal to 80, then the expected score of the game for the player with white would be 0.62 rather than just 0.50. The PCA rating system essentially uses this formula, with a value of C equal to 32 connoting a 32 rating point advantage for white.[38] This formula has strong connections to a model postulated by statisticians Roger Davidson and Robert Beaver in 1977.[39] Before a formula like that in (3) can be implemented, tournament data must be analyzed to estimate the value of C, and to substantiate or invalidate its adequacy and validity. For average tournament players, the advantage for white is less than it is for top players, so the value of C would be smaller than 80. This also suggests that the value of C might depend on the ratings of the players involved in a game.

[37] See section 8.93 of *The Ratings of Chessplayers*.

[38] The PCA determined this value by finding the average score for white in their database of over 100000 games. Their analysis, however, does not take the players' strengths into account, so it is likely that the true advantage conveyed to white is notably less than 32 points.

[39] "On extending the Bradley-Terry model to incorporate within-pair order effects," *Biometrics*, 33, pp 693–702.

Once an expected score formula that accounts for color is determined, the usual updating formula can be applied without modification based on these redefined expected scores. The main difference in updating from how the system currently proceeds is that players' ratings would not increase as much if they won with white, and would not decrease as much if they lost with black. Also, drawing a game against a higher rated player as white would earn fewer rating points than drawing as black. This reflects the knowledge that wins and draws are easier with white than with black.

Probability of a Drawn Game

The model we have used for describing the outcome of a chess game has assumed that only a win or a loss is possible. It is very curious, indeed, that the addition of a draw as a third possible outcome complicates the problem so greatly. Elo in his 1978 monograph dismisses the topic by arguing that information about the probability of drawing a game is not generally available.[40] It would probably be more accurate to say that the information is just as available as with measuring the expected outcome of the game, but incorporating draws into the rating system is much more difficult.

The simplest way to model the probability of a draw that relates to our model of values drawn from each player's box of numbers was described in a 1967 article by statisticians P. Rao and L. Kupper.[41] Their model assumes that a draw results when the values each player selects from their box are "close." This approach has some appeal because it implies that if two competitors exhibit roughly comparable playing strengths for a particular game, then the outcome of the game should be a draw. Rao and Kupper describe the procedure to estimate what constitutes closeness in playing strength. Suppose D is the largest difference in strengths displayed in a individual game that would result in a draw. Then Rao and Kupper show that the probability player A with true average strength R_A defeats player B with true average strength R_B can be expressed as

$$\Pr(A \text{ defeats } B) = \frac{10^{R_A/400}}{10^{R_A/400} + 10^{(D+R_B)/400}}. \tag{4}$$

The probability player B defeats player A can be computed by substituting R_A with R_B in the above formula (and vice versa). The probability of a draw can then be computed by subtracting these two probabilities from 1. A little bit of high school algebra shows that this formula implies that the probability of a draw is the same for any two players as long as the difference in their

[40]See Section 8.91 of *The Rating of Chessplayers.*

[41] "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," Journal of the American Statistical Association, 62, 194–204.

ratings is the same. Davidson and Beaver, besides describing how to incorporate the advantage to playing white in the Bradley-Terry model, also describe how to extend Rao and Kupper's model for drawn games to incorporate the advantage to playing white.

There are two major difficulties with this approach. One is that the model that leads to the formula in (4) may not actually be correct. At the very least, it might be reasonable to think that the frequency of draws would not only depend on the difference in average strengths of players involved in a game, but also the overall level of the players. For example, very strong players tend to draw games much more often than weaker players who are more prone to game-losing blunders. A second problem is that even if the formula is correct, it is not very clear how to use it to update ratings. One could compute an expected score of a game using the probabilities of a win, loss or draw, but no tangible advantage has been gained over the approach that is currently used.

Even though the system that is currently in place only calculates the expected outcome of a game, and that it is not directly connected to a simple probabilistic mechanism like randoming selecting numbers out of a box, it may be sufficient to describe playing strength. It may not be necessary to evaluate playing strength by modeling the probabilities of individual game outcomes. The tradeoff is that while potentially valuable information is lost when not modeling individual game probabilities, there is a realistic chance the model does not accurately describe frequency of game outcomes.

Incorporating the uncertainty of ratings

One feature of the rating system that has mostly been ignored, except in specific instances, is that some players' ratings are more poorly estimated than others. This can come about in two ways. First, players who have ratings based on the results of only a few tournament games are likely to have their abilities measured imprecisely. These players are treated by the rating system as provisionally rated, and their updating formula reflects the uncertainty in their ratings. Secondly, players who have not competed in tournaments for an extended amount of time have abilities that may be changing, so that their ratings become less reflective of their true average strength. The rating system currently makes no distinction among established players who compete regularly and those who compete sporadically. In both cases, alterations in the rating system updating procedure are required to incorporate the uncertainty in estimating ratings. Another instance in which uncertainty occurs is when an organizer is late in submitting a tournament report to the chess federation office. Administrators, such as the USCF, will rate events in the order they receive reports without regard to date of an event. Suppose two events, G and H, occur separated by 2 months with G occurring first. If the organizer for H submits a tournament report promptly, but the organizer for G waits, say, four months before submitting a tournament report, then H will be rated before G even though the two events occurred in reverse chronological order. This is a particular concern if a player has competed in both events. Under the current rating system, the earlier event (G) would in effect count more towards a player's current rating than the more recent event (H). It is clear that the results of the earlier tournament need to be downweighted relative to a more recent event, even if an organizer submits the report much later.

These problems can be reduced in several ways. One approach allows K in the updating formula to be a function of time since last competed and the number of tournament games played. As described earlier, K is a value that determines the amount of weight given to one's performance rating relative to one's pre-tournament rating. In the USCF system, once a player becomes established by competing in 20 games, K remains fixed at 32. The only exception to this rule is that when a player's rating is between 2100 and 2399, K becomes 24, and when a player's rating is at least 2400, K becomes 16. While the origin for this extension is not well-documented, one reason for its adoption is that players whose ratings are high are hypothesized to have abilities that do not change much over time, so K should be lower to reflect this stability. This may not, however, be a compelling reason to have K depend on one's rating.

When K is large, past performances are effectively downweighted relative to the current performance. Two instances where it might be useful to have a larger K than usual are when a player has a rating based on very few games so that the past performances are not precise indicators of ability, or when a player has not competed in a long time so that past performances may not be strongly indicative of current ability. The value of K should be low when a player is competing regularly so that his or her ability is well-represented by the player's pre-tournament rating. Also, K should be low when an organizer has submitted a tournament report much later than the tournament's ending date when more recent performances have been rated. For example, if a tournament was completed in June 1992, and the results were submitted in August 1993, and a player's current rating is based on results from a tournament in July 1993, then the results of the competition in June 1992 may have little bearing on the player's current strength relative to the event in July 1993. When the player's rating is updated, little weight should be given to this performance from a year earlier.

In addition to changing K in the updating formula depending on the uncertainty in a player's pre-event rating, a similar modification is necessary when updating such a player's opponent's rating. For example, if an established player rated 1700 is defeated by another established player rated 1700, the first player's rating decreases 16 points. If the second player had a provisional rating of 1700 based on only having played 4 tournament games, and the established player is defeated, then the current system again says the established player should lose 16 points. But in this second situation the player whose rating is provisional is possibly a much better player who had a poorly estimated rating, so the established player should not lose as many rating points. Thus a player who competes against an opponent whose K is large should have that game result in a fraction of rating points gained or lost.

This approach suggests that K should vary for an individual from game to game. The formula to compute K for an individual game could be constructed through statistical modeling. Having obtained the value of K_{game} for each game in a tournament, a player's rating would then be updated by summing

$$K_{game}(S_{game} - S_{game.exp})$$

for each game, where S_{game} is the score of a game (0, $\frac{1}{2}$ or 1), and $S_{game.exp}$ is the expected score for the game based on the expected score formula. The resulting computation would more appropriately take into account the uncertainty contained in ratings.

A more formal approach to incorporating uncertainty into the rating system is to describe knowledge about a player's unknown rating parameter not simply by an estimate, but by both an estimate *and* a measure of variability of this estimate. This measure of variability describes how much faith one should have in the rating estimate. For players who have only played a few tournament games or who have not competed in a long time, the variability measure associated with the rating estimate will likely be large. Players who compete regularly will have measures of variability that are small so that their ratings are reasonably indicative of their rating parameters. The measure of variability, in conjunction with the rating estimate, can be used to provide a *range* of likely values that a player's rating parameter takes on. Instead of just reporting a "best guess" of a player's rating parameter, as the currently implemented system attempts to do, this extension can give a plausible interval of values of the rating parameter; the interval is wider for players whose rating estimates are more uncertain.

As with the approach based on varying K, the differing measures of variability from player to player have consequences on the magnitude of rating changes. For instance, when one player has a rating with a large associated variability (indicating that the player's rating is an imprecise estimate of his or her rating parameter) and an opponent's rating has low variability (indicating the opponent's rating is precise), then the results of the game should have a large impact on the rating of the player whose rating has large variability, but only a modest effect on the rating of the opponent's rating. Another similarity with the approach based on varying K is that the passage of time has an effect on the variability of one's rating estimate. As time passes, the measure of variability can be increased to reflect the extra uncertainty in one's ability. In fact, the system can be modeled so that certain players, such as younger players, can be assumed to have measures of variability that increase more quickly over time than adult players, whose abilities likely do not change as much. The main advantage to this more formal approach than with the varying K approach is that the expected score function can be changed to incorporate the measures of variability. Specifically, the expected score of a game played between two players with uncertain rating estimates is closer to 50% than the usual formula predicts – this argument was used earlier to describe the reason the dotted line in Figure 6 did not intersect the segments. The computation of the expected score incorporating the measures of variability can be derived using integral calculus, but approximated numerically by a simple formula.[42]

It should be noted that one of the consequences of incorporating uncertainty of rating estimates into the rating system is that the rating gain for one player need not be the rating loss for the other. The magnitude of the changes would depend on the variability of each player's rating. This might seem, at first, to violate some underlying principle that rating points in the rating system must be conserved, but this principle is really a myth. No technical or theoretical reason necessitates that rating points must be conserved. It is this principle that is partly responsible, as argued earlier, for rating deflation. Appropriately incorporating measures of variability about rating estimates is one method to tackle the problem of deflation.

Competing incentives

One of the most important problems with the rating system has little to do with its computational aspects or the validity of its assumptions; it has to do with players' perceptions of ratings, and the consequences of these perceptions. While the implementation of a chess rating system has increased the popularity of tournament chess, it may also be equally responsible for those who leave tournament chess.

The rating system has become equated with a reward/punishment system. Even the terminology associated with ratings demonstrates this. When a player's rating increases, the player is often said

[42]The details of the calculations are found in "An Extension of the Elo Rating system," an unpublished paper by the author.

to have "gained" rating points, and a player's rating decreasing corresponds to rating points "lost." So a player who loses games in a tournament must accept the additional insult of losing rating points as well. This interpretation of ratings may cause discouragement among players whose ratings continue to decline, and subsequently cause them to refrain from tournament participation for fear of losing more rating points. The view that declining ratings are a punishment or insult is a disincentive for players to compete. This view of ratings need not be inevitable because a lowering of one's rating may merely indicate that a player was initially overrated, not that a player's ability is becoming worse. But issue remains that a rating change may be perceived as a reflection on one's pride or self-esteem rather than as an improved estimate of one's ability.

This notion of a reward/punishment system is further enhanced by the construction of rating "classes" that correspond to rating ranges. For example, if a player's USCF rating falls between 1800 and 1999, the player is called a "Class A" player; if the rating falls between 2000 and 2199, the player is called an "expert"; if the rating falls between 2200 and 2399, the player is called a "master." When a player's rating crosses a boundary that places him or her in a higher class, a sense of achievement has been garnered. Similarly, when a rating drops below a class boundary, disappointment may result.

Even more consequential is that tournament organizers in the U.S. section tournaments according to rating classes. Not only does this systematically preclude players whose ratings are just above a rating class boundary due to the imprecision of their rating from participating in an appropriate class section, but it also provides an incentive for players to manipulate their ratings by artificially lowering their ratings. This is accomplished by purposely losing games in unimportant tournaments. The current design for organizing tournament sections and the reward/punishment interpretation of ratings make it difficult to view ratings simply as a means to measure ability and predict future game outcomes.

Fortunately, in the last few years, an additional system has been developed called the "title" system. This system is intended to complement the current rating system by functioning as a reward system. At the August 1993 delegates meeting, an overwhelming number of organizers even agreed that they would experiment by sectioning their tournaments according to titles rather than according to ratings. The system does not intend to track players' abilities as the rating system is designed to do, but instead it awards players for incremental improvements in their performances.

The USCF title system is based on the philosophy that an exceptional tournament performance should be rewarded, but a poor tournament performance should simply be ignored. To earn an

"1800-title", a player would need to achieve performances in tournaments that would provide strong evidence that the player is of at least 1800 strength. Under the current system, such a player would need to demonstrates five extraordinary performances, or "norms," in order to acquire the title. If a player has accumulated four norms toward the "1800-title" and has a poor result in a subsequent tournament, this would have no effect on his four accumulated norms. This system only rewards positive results and does not punish poor results.

One of the crucial aspects of the USCF title system is that acquiring norms is completely independent of one's own rating, though it does depend on opponents' ratings. The same norm is awarded to a player with a high rating as one with a low rating if they attain the same score against the same opponents. This is an important idea because it lessens reliance on one's own rating as a measure of chess achievement, which a rating was not intended to be in the Elo system.

The USCF title system has strong connections to the system used by FIDE for awarding titles, such as titles of grandmaster and international master. In this system, players must achieve outstanding results in events with very highly rated players in order to acquire norms. The higher the average FIDE rating of players in an event, the lower the score needed to obtain a norm. As with the USCF title system, norms are never lost due to poor results. The USCF title system also has strong connections to the ACBL bridge rating system which awards master points only to positive performances, and never subtracts points for poor performances.

A direction that would relieve the rating system of the burden of functioning as a reward/punishment system would be to emphasize titles as the object of attainment, not an increased rating. Class designations of ratings need to be stripped away and associated solely with titles to restore the unconfounded interpretation of ratings as measures of ability. The less attention players place on their rating, the less reason players will feel discouraged by rating decreases. Furthermore, titles provides players with an incentive to continue participating in tournaments without the risk of the losses associated with rating fluctuations. As Macon Shibut, editor of *Virginia Chess*, has argued in an unpublished article, the titles system needs to become part of chess culture much in the same way that the current rating system has become.[43]

[43] "USCF Lifetime Titles; A Good Idea, But Will It Fly?" 1993

Conclusions

The Elo rating system is based on two simple formulas – the formula that describes the expected score of a game given two players' ratings, and the formula that describes how a player's rating changes over time. As this article has described, assumptions are required to apply these formulas, and rethinking these assumptions may result in the need to modify the current formulas so that ratings have sensible interpretations.

When the USCF rating system was implemented in the early 1960's, players' ratings were kept on index cards and rating updates were computed by hand. As membership grew and tournament frequency increased, updating ratings by hand had become a tedious task. Carrying out such a procedure today would be unthinkable. With more than 30000 USCF members playing every year, and thousands of tournaments organized every year, the USCF relies on the power of computers to perform rating computations, as well as a variety of other membership-based functions. Fortunately, because ratings are updated through the computer, modifications in the algorithm are not hindered by the complexity of the changes. As the assumptions underlying the rating system are continually questioned and tested, changes in the rating algorithm can reflect our understanding of the frequency players win chess games and how players' abilities improve over time.

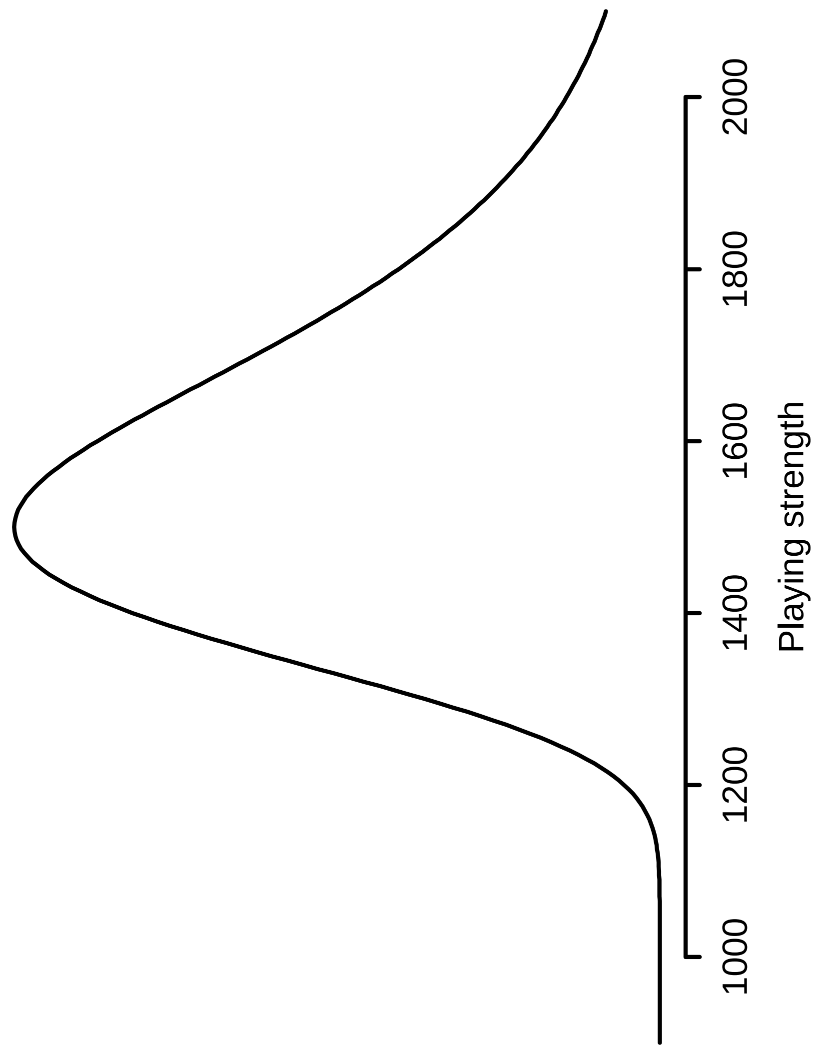

Figure 1: An extreme value distribution centered at a strength of 1500. Higher points on the curve indicate greater likelihood that a player will perform at that level.

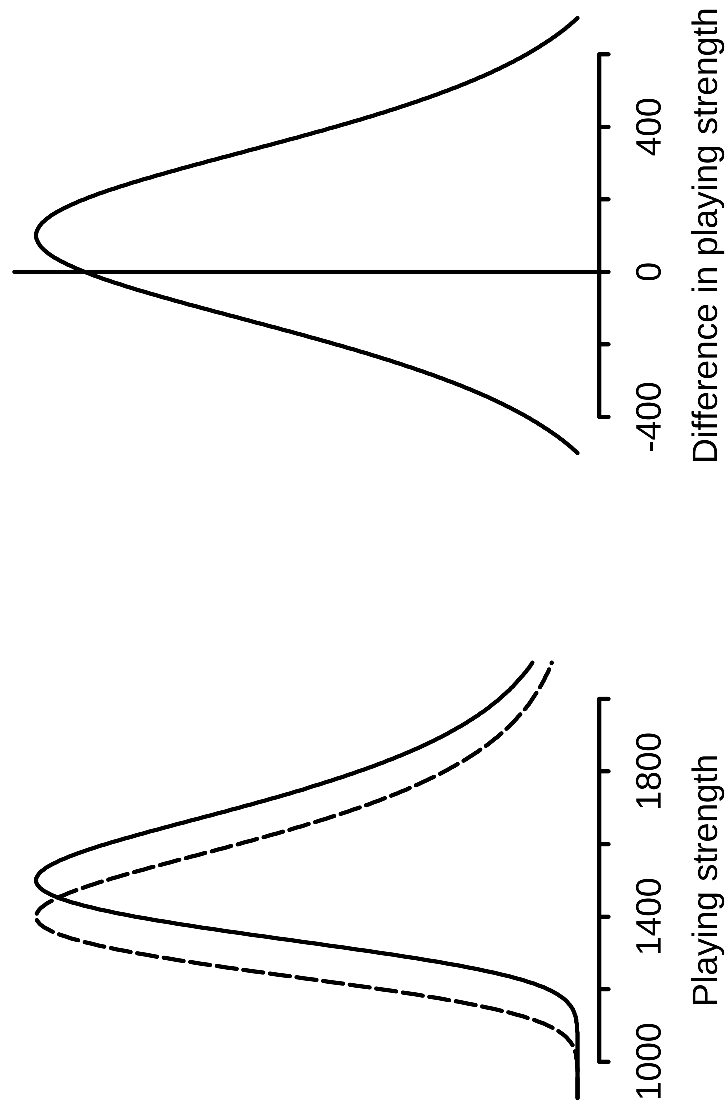

Figure 2: *Left*: Two superimposed extreme value distributions, one centered at 1400 (dotted line) and one centered at 1500 (solid line). *Right*: Logistic distribution of the difference between two players' individual performances. The area under this curve to the right of 0 is the probability the stronger player will outperform the weaker one.

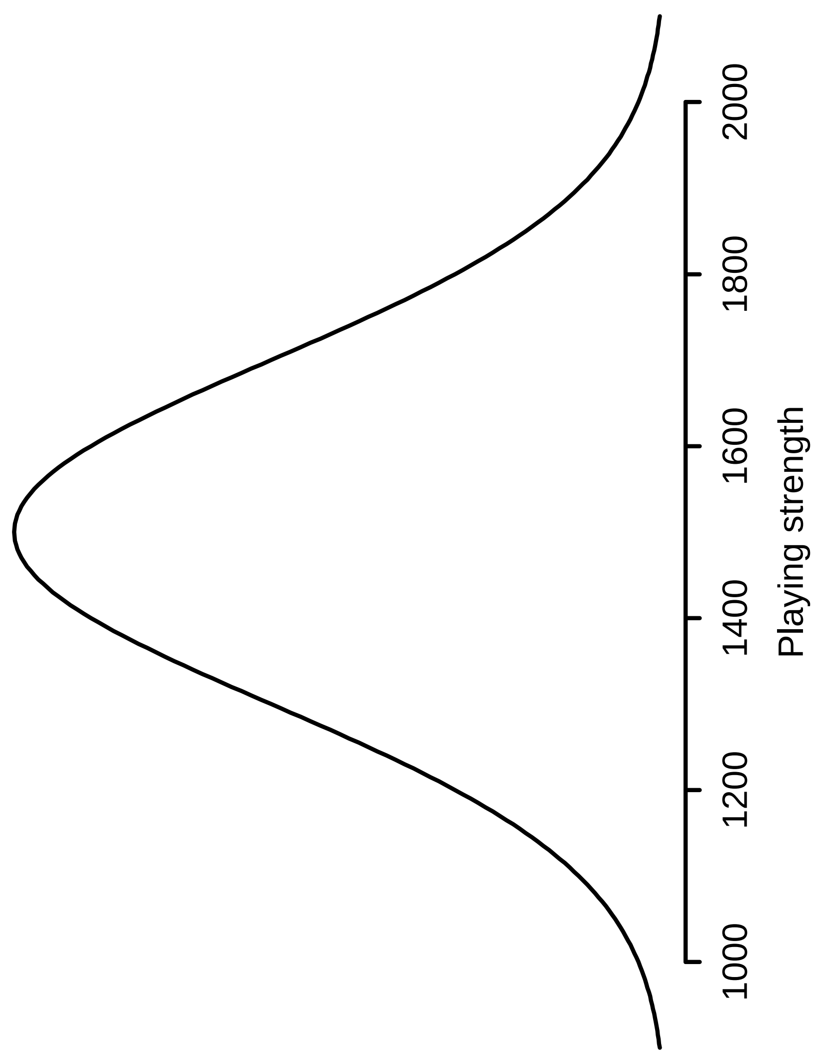

Figure 3: Normal distribution centered at 1500. As in Figure 1, higher points on the curve indicate greater likelihood that a player will perform at that level.

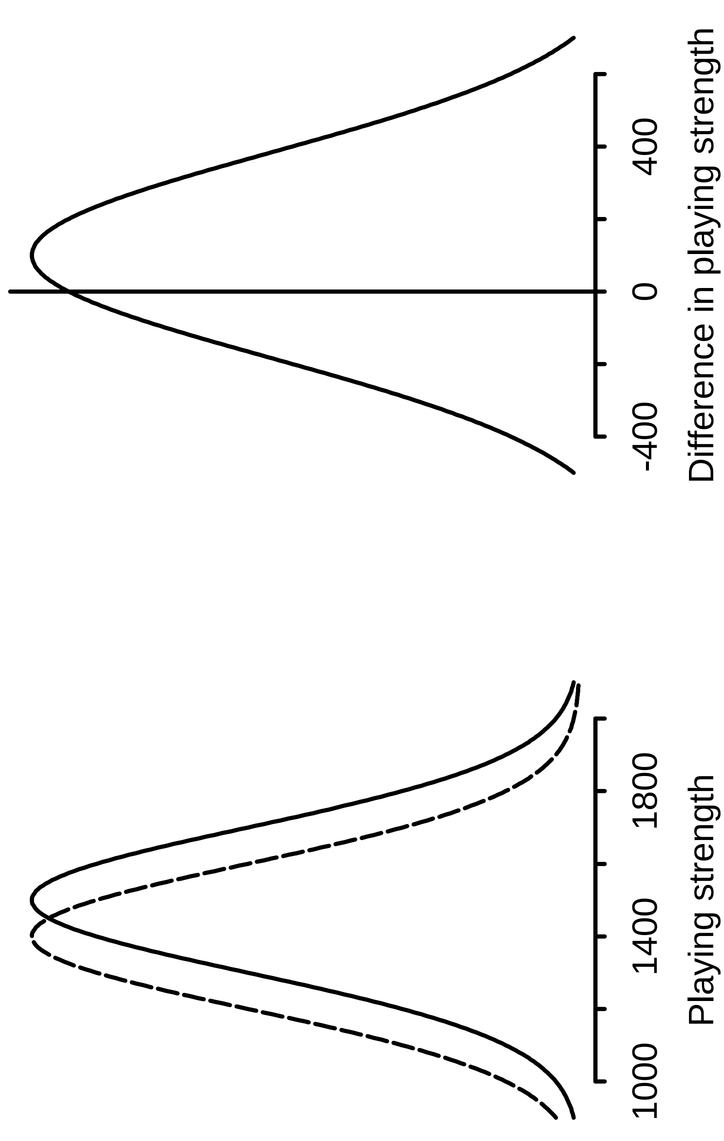

Figure 4: *Left*: Two superimposed normal distributions, one centered at 1400 (dotted line) and one centered at 1500 (solid line). *Right*: Normal distribution of the difference between two players' individual performances. The area under this curve to the right of 0 is the probability that the stronger player will outperform the weaker one.

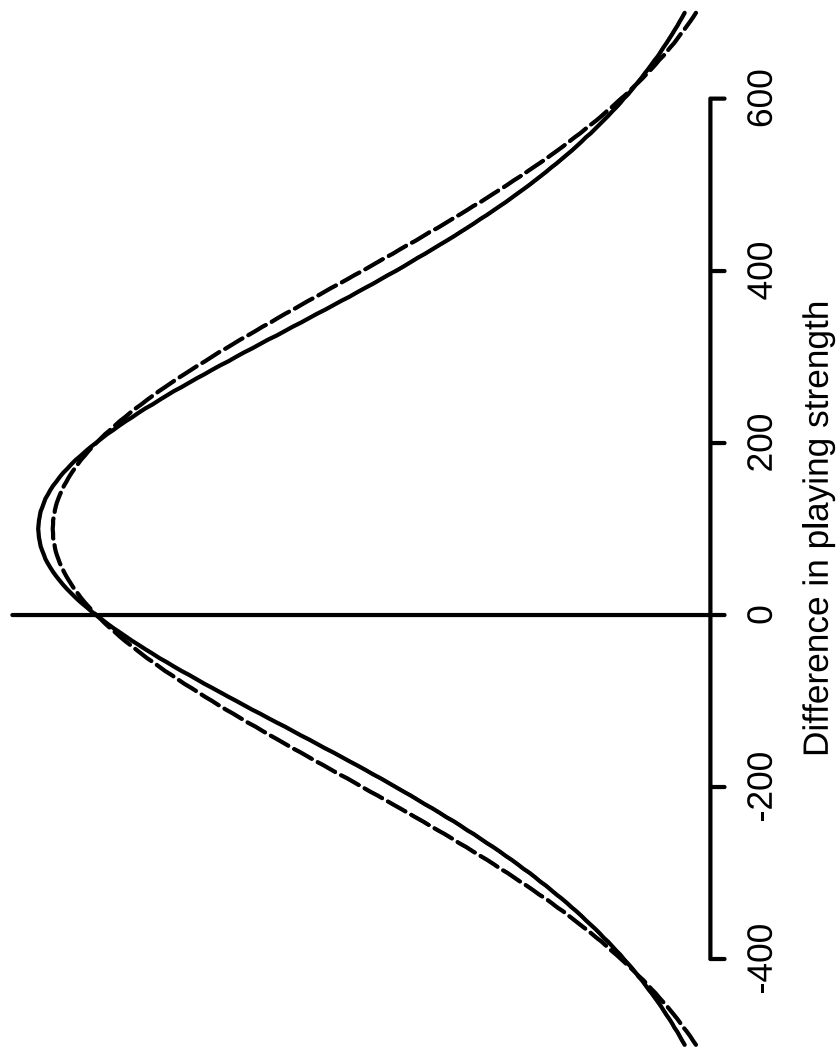

Figure 5: Two superimposed distributions of the difference between two players' performances – the logistic distribution (solid line) and the normal distribution (dotted line). For practical purposes, the two curves are indistinguishable.

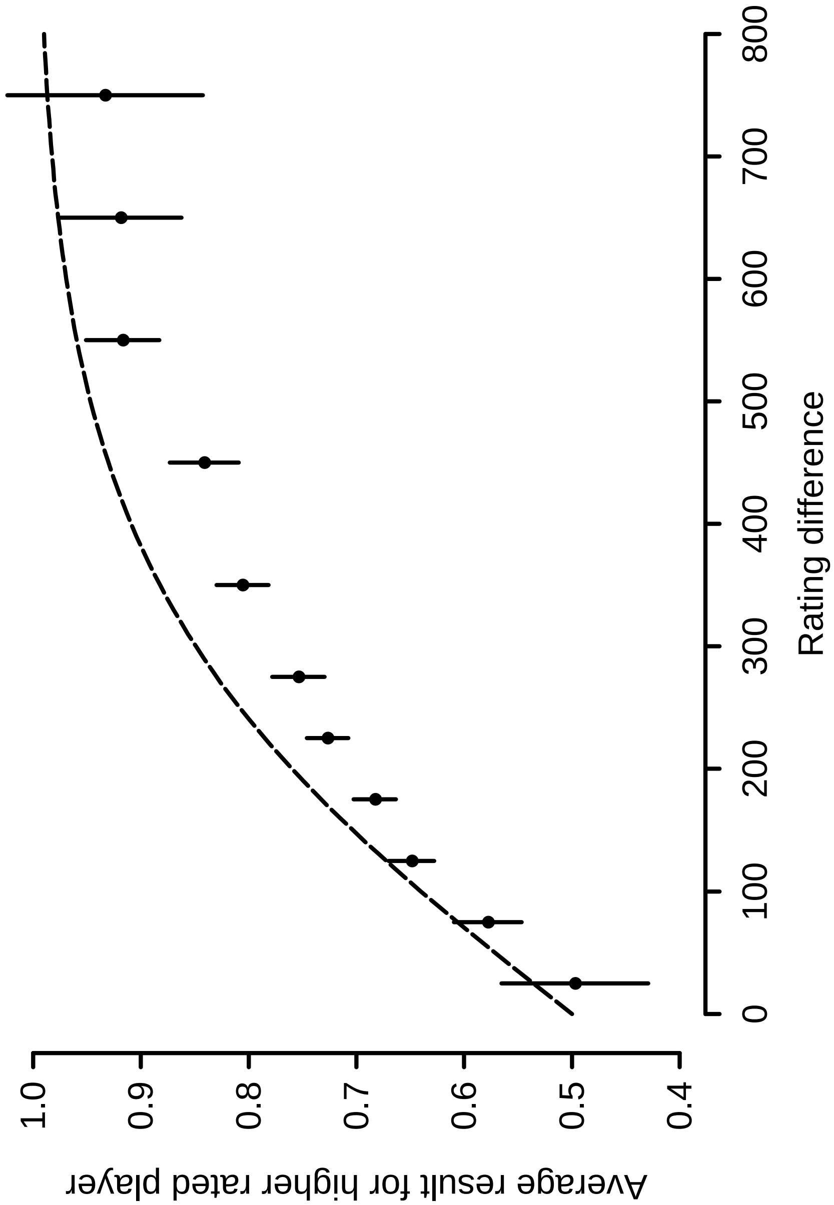

Figure 6: Summary of 8329 rated USCF tournament games. Both players must have competed in at least 20 tournament games to be included in the sample. The sample is partitioned into groups of players according to their rating difference (0–50, 50–100, 100–150, 150–200, 200–250, 250–300, 300–400, 400–500, 500–600, 600–700, 700–800). For each rating difference group, the dot represents the average score of games relative to the higher rated player. The vertical bars show the 95% margin of error. The values on the dotted line are the expected scores calculated from Elo's expected score formula.

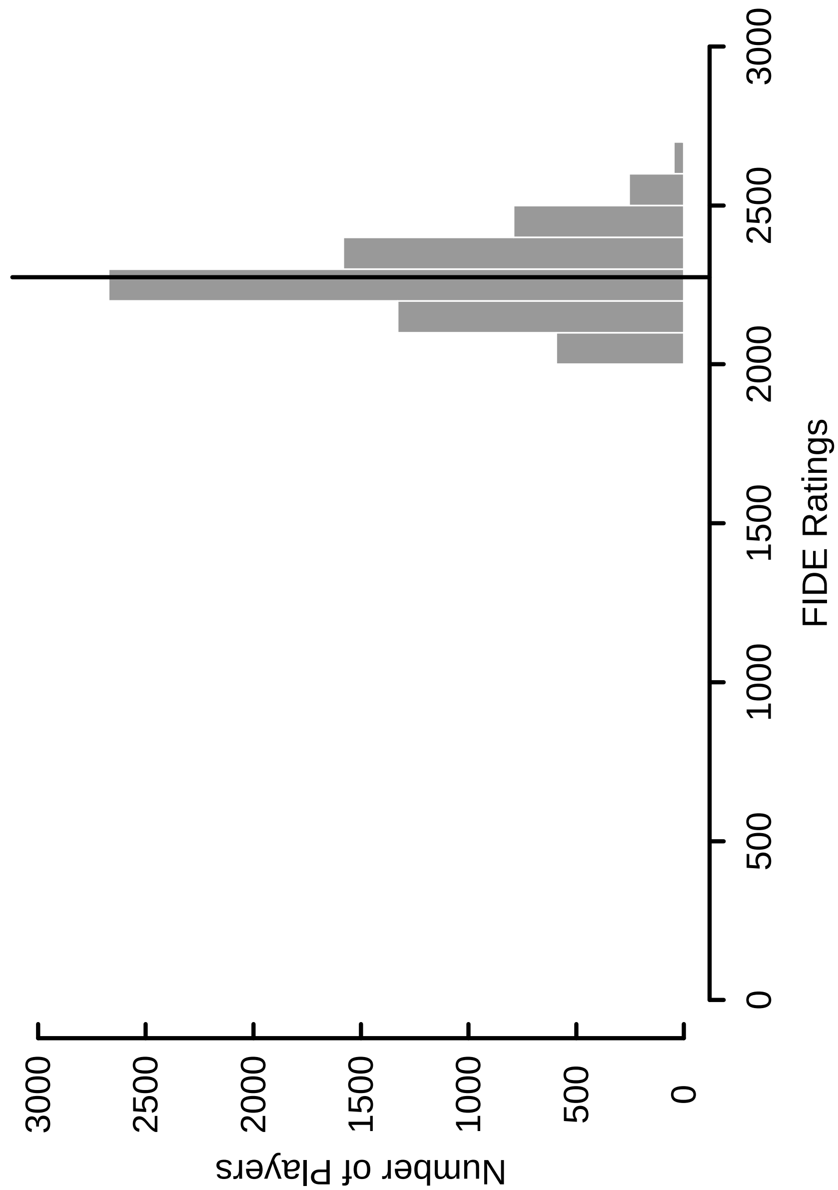

Figure 7: Distribution of FIDE ratings, July 1994. Players who competed in at least one FIDE-rated game in the previous six months are included in the sample.

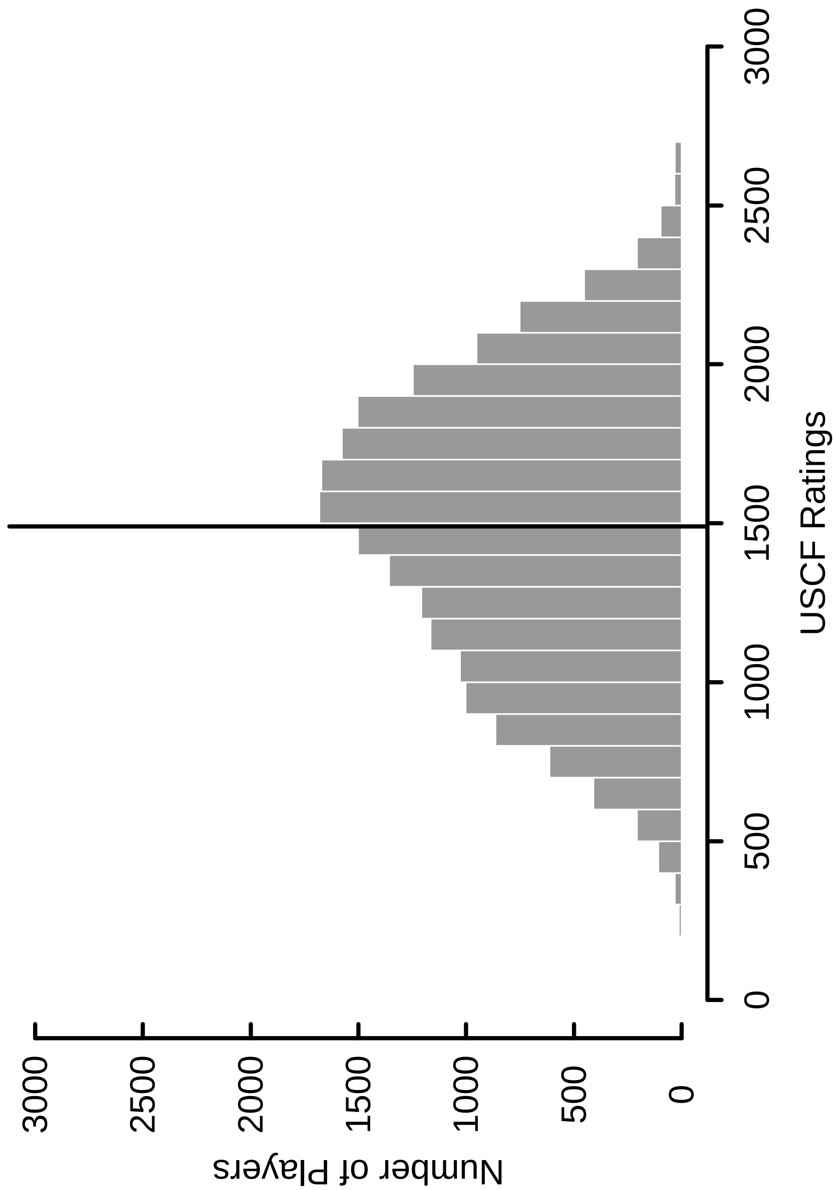

Figure 8: Distribution of USCF established ratings, July 1994. Players who competed in at least one USCF-rated game in the previous six months are included in the sample.

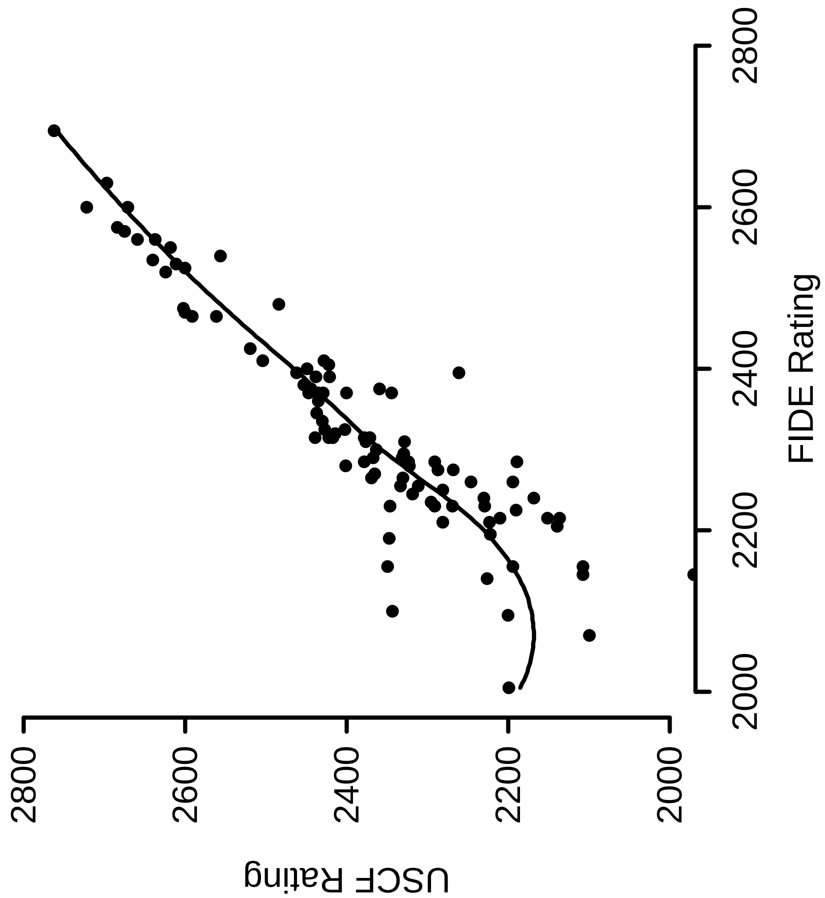

Figure 9: Plot of USCF ratings against FIDE ratings for 211 players common to both July 1994 rating lists. Players who played at least six FIDE-rated games in the previous six months and at least one USCF-rated game in the previous six months, and had achieved an established USCF rating, were included in the sample. The curve that traces through the data is a "locally weighted scatterplot smoother" which summarizes the relationship between USCF and FIDE ratings.

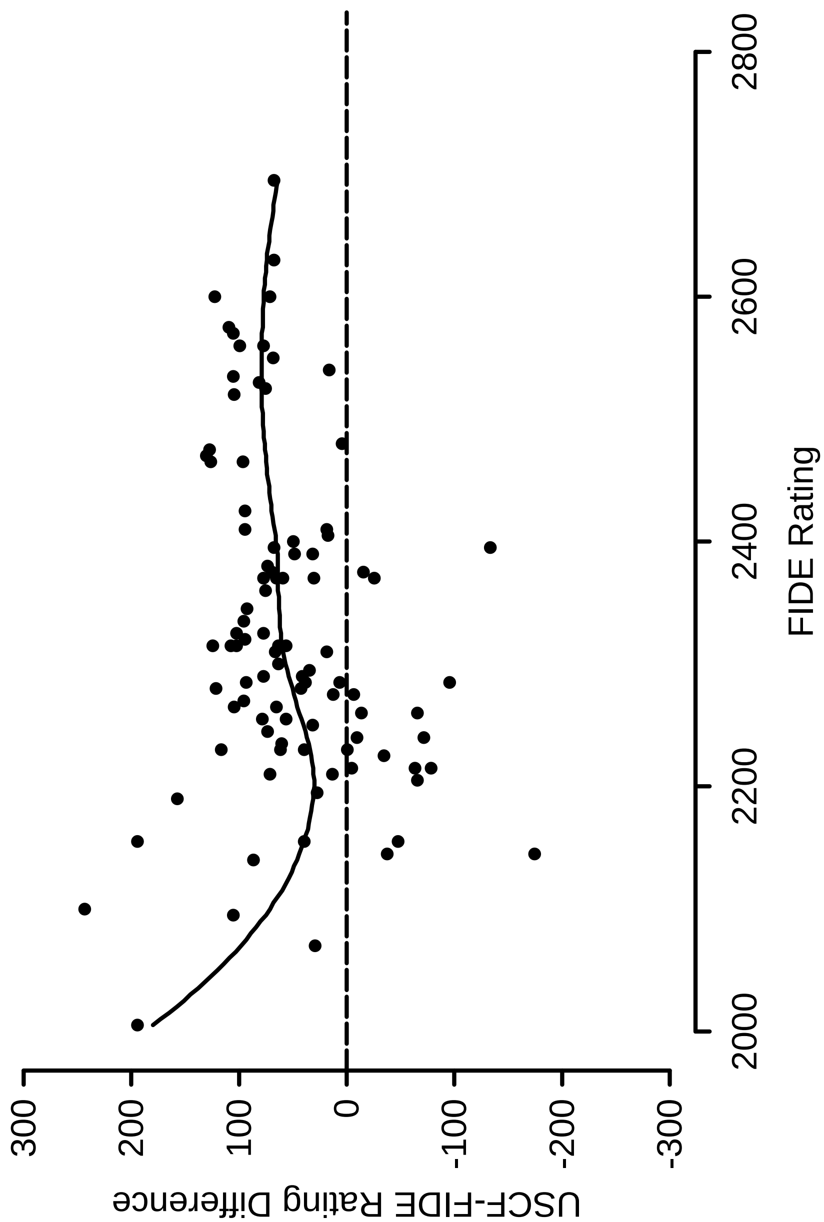

Figure 10: Plot of USCF–FIDE rating difference against FIDE ratings for 211 players common to both July 1994 rating lists. The criteria for inclusion in the sample are the same as those for Figure 9. The scatterplot smoother demonstrates that the USCF–FIDE average rating difference depends on a player's FIDE rating. For players with a FIDE rating of 2050, the expected USCF–FIDE difference is 120; for players with a FIDE rating of 2200, the expected difference is 30; for players with a FIDE rating of 2550, the average difference is 70.