Rating the Chess Rating System

Mark E. Glickman[*]
Department of Mathematics
Boston University
mg@math.bu.edu

Albyn C. Jones
Department of Mathematics
Reed College
jones@reed.edu

The introduction of chess rating systems may have done more to popularize tournament chess than any other single factor. In the 1950s, Arpad Elo (1903–1992) developed the theory of the current U.S. rating system, often called the "Elo system." Elo based his scale on one previously used by the U.S. Chess Federation (USCF), which was calibrated relative to the performance of an "average" player in a U.S. Open Championship. Elo's system, however, added considerable statistical sophistication. Since its development, 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 many 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' 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 Elo rating system calculates for 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 on the outcomes of tournament games. When two players compete, the rating system

[*] We wish to thank Christopher Chabris for his helpful discussions.

predicts that the one with the higher rating is expected to win more often. The more marked the difference in ratings, the greater the probability 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. 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 to permit such a comparison. But do they really predict game results accurately?

A model for chess strength

While Elo's name is by far the most associated with the development of the current chess rating system in the U.S., the statistical principles underlying the system had been established well before his work in the late 1950's, and certainly prior to his well-known 1978 monograph (Elo, 1978). As far back as the late 1920s, Zermelo (1929) addressed the problem of estimating the strengths of chess players in an uncompleted round robin tournament. Good (1955) developed a system that amounted to the same model as Zermelo's, but was obtained through a different set of assumptions. Both of their models are connected to the Bradley-Terry model for paired comparison data (Bradley and Terry, 1952). Among standard paired comparison models, the Bradley-Terry model has the strongest connection to the USCF's implementation of the Elo rating system.

The fundamental assumption of Elo's rating system is that each player possesses a current playing strength, which is unknown, and that this strength is estimated by a rating. In a game played between players with (unknown) strengths of R_A and R_B, the expected score of the game for player A is assumed to be

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

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. Suppose, for example, that the strengths for players A and B are 1500 and 1700, respectively. Then the above formula states that the long-run average score for A is about 0.24. As a statistical model, the formula in (1), which is often termed the "winning expectancy" formula, applies only to the unknown playing strengths. As we describe below, the formula is also used in the rating procedure by replacing the parameters with their estimates. The main difference between this model and the Bradley-Terry model is that the Bradley-Terry model only applies when the outcomes are binomial (a win and a loss), so that the formula in (1) is a probability rather than an expected score.

Updating ratings

Suppose a chess player has an established USCF rating prior to a tournament. An established USCF rating is one that is based on at least 20 tournament games. The rating update formula involves adjusting a player's pre-tournament rating with the observed results in the tournament. The adjustment is made based only on the current tournament, so that rather than recomputing a 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 can be thought of as averaging an estimate of the playing strength demonstrated in the tournament with the pre-tournament rating. 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 rating, r_{pre} is a player's 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 estimated winning expectancies, E, for each game using formula (1).

The above formula has some interesting 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 achieved a result better than the pre-tournament rating predicted, so the player's rating is increased to reflect the possible improvement in strength. Similarly, if the term $(S - S_{exp})$ is negative, then the player performed worse than expected, and this player's rating will decrease by the discrepancy magnified by the value K. The larger the discrepancy, $(S - S_{exp})$, in magnitude, the more we are inclined to doubt the pre-tournament rating, and thus the greater the change required to 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 also seems appropriate to note that even if a player is correctly rated, random variation in the results will produce variation in the rating, which will be naturally corrected in the long run.

The attenuation factor K in formula (2) can best be interpreted as the amount of weight given to the new tournament results relative to the pre-tournament rating. The larger the value of K, the greater the amount of change allowed in one's rating. In the current USCF rating system, K is either 32, 24 or 16 depending on the pre-tournament rating. If the player has a very high tournament rating (2400 or higher), $K = 16$; slightly lower ratings ($2100 \leq r_{pre} < 2400$) correspond to $K = 24$, and the rest have $K = 32$. The rationale is that stronger players tend to have more stable abilities, so their ratings are not expected to change much.

An analogy can be drawn between the formula in (2) and tracking the position of a moving object. Suppose we have a rough idea about the current location of an object. The object now moves, and our tracking instrument tells us its approximate location. The tracking instrument will likely assume that the position of the object cannot be too far from the previous location, so that updating the measurement will be based on information of the prior measurement. This is analogous to tracking a player's chess ability. A player's pre-tournament rating roughly conveys current playing strength, or the player's current "position." The expected score formula summed across opponents corresponds to a prediction based on the object's last known position. An actual total score is observed, and we adjust our estimate of the player's current "position" by using the formula in (2). The rating system can therefore be viewed as a device that tracks a player's ability as it changes.

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 is not used when a player has no rating prior to entering a tournament. The USCF has implemented a system to compute initial ratings using a different set of formulas. The resulting estimated ratings are called “provisional ratings.” As the name implies, we do not attach a great amount of confidence in provisional ratings because they are estimates based on a very small sample of game outcomes.

When a player has competed in fewer than 20 tournament games, the post-tournament rating is calculated based on all previous games, not just the ones in the current tournament. The provisional rating formula is

$$r_{post} = \bar{r}_{opp} + \frac{400(W - L)}{N} \tag{3}$$

where r_{post} is the player’s post-tournament rating, $\bar{r}_{opp}$ is the average of the opponents’ ratings, W is the number of wins, L is the number of losses, and N is the total number of games.

In actuality, the provisional rating computation is more complicated than the formula in (3). For example, if an unrated player competes against another unrated player, the formula cannot be used. This is addressed in the USCF rating system by imputing a rating of 1000 for the unrated opponent. An undesirable consequence of the provisional rating formula is that a player’s rating can decline as a result of winning a game. Suppose a player has scored a win, a draw and a loss against opponents rated 1400, 1500 and 1600 in his last tournament, giving him a provisional rating of 1500 based on 3 games. He now defeats a player rated 700. According to the formula, his rating after the fourth game is now

$$1300 + \frac{400(2 - 1)}{4} = 1400$$

which is a drop of 100 points. Winning a game is clearly never evidence that the player is overrated. This problem is addressed by adding a condition to the provisional rating calculations that prevents a rating from declining based on a win.

Does the USCF rating system work?

The method Elo laid out for adjusting ratings was adopted by the USCF in 1960. Over the intervening years various modifications have been introduced to the USCF system to deal with perceived problems. It is natural to ask, therefore, whether the current system produces ratings that predict performance accurately. We first examine the distribution of USCF ratings before addressing this question.

The upper histogram of Figure 1 shows the distribution of players with established ratings in January 1998. The mean rating for established USCF players in January 1998 is 1337. USCF established ratings ranged from 0 to 2751. About 96.5% of all USCF established players have ratings less than 2200, the rating at which a player is considered to be a "master."

The histogram also indicates that the distribution of player strengths is bimodal, with peaks in the 700-800 range, and in the 1500-1600 range. This can be explained by the large number of scholastic (K–12) tournament players. As seen in the lower histogram in Figure 1, there are a large number of established tournament competitors between 10–15 years old. In fact, this group accounts for more than 25% of all established tournament players.

The relationship between players' ages and ratings can be seen in Figure 2. The plot consists of all players with established ratings in January 1998 for which age information was recorded. Out of the 27563 established players, only 1787 had missing age information. For players under 20 years old, the plot indicates ratings centered near 1000, while for older players the average rating is closer to 1500. In general, younger players' ratings tend to increase with respect to age until they are about 35–40 years old. Then ratings level off near 40 years old, and finally decline gradually in the later years.

Can the winning expectancy formula be used to predict game outcomes between pairs of established players? To answer this question, we examined the outcomes of all tournament games between players with established ratings played between January and October 1997. There were 225,621 games in this sample. If the winning expectancy formula is accurate, then we should expect that, for fixed rating difference ΔR, the average score of a large collection of games with rating difference ΔR should be close to the expected score given by the winning expectancy formula, $1/(1 + 10^{-\Delta R/400})$. Figure 3 shows the results of our 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 95% confidence intervals. The dotted line in the figure corresponds to the winning expectancy in (1). If ratings were predictive of game outcomes, then the dotted line would intersect the segments on the figure. With very few exceptions, the confidence intervals computed from the observed data underestimate the theoretical winning expectancy. Thus, lower-rated players are scoring better than predicted by the ratings and the model, and that this behavior is consistent across all rating differences.

Based on the poor fit to the winning expectancy formula, we postulated a model for the average observed score as $1/(1 + 10^{-\alpha(\Delta r/400)})$, where Δr is the observed rating difference, and α is an unknown scaling parameter to be inferred from data. The maximum likelihood estimate of α

(treating all the games as independent given the ratings) was $\hat{\alpha} = 0.713$. This corresponds to replacing 400 in the winning expectancy formula with 561, so that a more useful prediction of a game outcome based on published ratings is $1/(1 + 10^{-(\Delta r/561)})$. The fitted model is drawn on Figure 3 as a dashed line. It is quite remarkable how well this model fits to the data, which can be seen by how consistently the dashed line traces through the confidence intervals.

What is going on here? Why is the winning expectancy formula's prediction too high for the higher rated player? One explanation is that lower-rated players tend to improve more quickly than higher rated players, so while a rating is "official" a lower-rated player may already be substantially better than the official rating indicates. This would suggest that the true (unknown) difference in strengths is smaller than the published ratings indicate.

A more subtle explanation for the formula's over-optimism of the higher rated player is to consider that ratings are merely estimates of playing strength, and are therefore subject to variability. To consider an extreme example, suppose two players have published ratings that differ by 400, but that neither player has competed in many years so that their ratings are practically meaningless. This would be a situation where the variability of the rating estimates are extremely large. Intuitively, because the ratings do not reflect the players' current strengths, the expected outcome between the two players should be close to 0.5 even though the ratings predict a 0.91 winning expectancy for the higher rated player. Thus, in this extreme case, the winning expectancy formula over-predicts for the higher-rated player.

Even if the ratings are more precise estimates of strength, the variation in the estimate can still account for a smaller winning expectancy for the higher-rated player. Consider the following crude example: A player with a strength of 1900 plays against an opponent with a reported rating of 1700. Suppose that half of all players rated 1700 were really 1600 strength, and the other half were 1800 strength. If we calculate the winning expectancy using the opponent's reported rating of 1700, we obtain a value of 0.76. In truth, the winning expectancy is 0.64 if the opponent is 1800 strength, and is 0.85 if the opponent is 1600 strength. So, on average, the first player can expect to score

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

against the opponent. This value is less than 0.76, which is the result computed on the reported rating of 1700. In general, averaging the winning expectancy over the uncertainty in the ratings produces a lower value (i.e., closer to 0.5) for the higher rated player.

Because $0.713 = \hat{\alpha} < 1$, we can conclude there is a fair amount of variability in rating estimates.

Rating range for average rating	Number of Games	Estimate of α
1 – 500	1383	0.946
500+ – 800	14366	0.919
800+ – 1200	41148	0.805
1200+ – 1400	25649	0.634
1400+ – 1600	36507	0.590
1600+ – 1800	40059	0.653
1800+ – 2000	33982	0.733
2000+ – 2200	21261	0.829
2200+ – 2700	11252	0.950

Table 1: Estimated values of α by level of competition. Each game in the sample of 225,621 was stratified into one of nine groups according to the average rating of the two players involved in a game. Smaller values of α indicate greater uncertainty in players' published ratings.

In fact, the smaller the value of α, estimated from the data, the more variability we can attribute to the published ratings. One issue we decided to explore is whether the variability of ratings depends on the strength of the players. To answer this, we divided our data into nine groups according to the average rating within a pair of players, and determined the maximum likelihood estimate of α within each group. Table 1 displays the estimated values of α. Because higher-rated players tend to have more stable abilities than lower-rated players, and that they tend to compete more frequently, it is not surprising that the estimated value of α tends to be close to 1. This can be seen from the higher rating groups where the estimates of α are larger than 0.8. Starting from the strong players, the estimated values of α decline as rating level declines. This trend continues down to the 1400–1600 range. But, surprisingly, the trend reverses for the low end of the scale at which point α becomes large again. We expected that the estimated values of α for the low-rated players would be even smaller than the middle group because we thought that many low-rated players, most of whom are scholastic players, would have a tendency to improve rapidly, so that their published ratings would not be precise estimates of current strength. One possible explanation is that early in the development of cognitive expertise, a person may remain stagnant before undergoing growth. What we may be seeing is that many of these low-rated chess players are at this stagnant point in development, so that their abilities, which are temporarily stable, are well estimated. This hypothesis is an area for future investigation.

Apart from the differing amounts of variability at different levels, another source of variation in game outcomes is the advantage of having the first move. In chess, the player having the white pieces moves first, and it is well understood that playing white conveys a tangible advantage. Elo (1978) estimates that among players with similar strengths, the expected score for a player with the first move is 0.57, which corresponds roughly to a 50 point rating difference. Elo's model does not recognize this advantage, so not incorporating color assignment can be viewed as another source of variability. Unfortunately, tournament directors are not instructed to keep or report color assignment for each game, so this information is not available.

Can the imprecision of ratings and not knowing color assignment explain the values of α that "correct" the winning expectancy formula? Using an approximation explained in Glickman (1998), the standard error of an established rating required to account for the overall factor of 0.713 is 220. For the higher rated players, the factor of 0.95 corresponds to a standard error of 70, and for the middle group the factor of 0.590 corresponds to a standard error of 300. Thus, for a player rated 1500, an approximate 95% confidence interval of the player's true strength is (900, 2100). Most tournament players would likely argue that ratings are more precise estimates than such a large standard error implies, even without knowledge of color assignment, so we are left with the inescapable conclusion that it is more than just the imprecision in ratings that explains the phenomenon seen in Figure 3. This is still the conclusion reached even accounting for the uncertainty due to color assignment, which reduces the standard errors by only a small amount.

We now discuss some of the challenges that make the task of rating chess players difficult.

Isolated rating pools

The title of the recent play by John Guare, "Six degrees of separation" (1990), 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.

In measuring chess ability, this notion of being able to trace paths that connect players has direct relevance. While we will not claim 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 reasonably accurate predictors of how each will fare against the other, assuming one trusts the winning expectancy 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 through their ratings that the players in the first group are better players on average than those in the second group. A player rated 1500 in the first group will likely be notably stronger 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 tend to compete only 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 when they do, they rarely return exclusively 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.

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 is to lower the overall average rating.

In the mid-1970's, it was becoming apparent that the average rating of USCF players was beginning to decline. 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. If, for example, a player's highest attained rating was 1871, then the player's rating could not decline below 1600. 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 unlimited

Rating Status		January 1997	January 1998	Mean rating	Number of
1997	1998	Mean	Mean	increase	Players
Established	Established	1430	1440	10	18498
Established	Inactive	1310	—	—	9094
Provisional	Established	910	940	30	3961
Provisional	Provisional	890	900	10	3862
Provisional	Inactive	860	—	—	14051
Inactive	Established	—	1280	—	5104
Inactive	Provisional	—	820	—	18288

Table 2: USCF Rating summaries for January 1997 and January 1998. A player's rating is either established or provisional in January if the player competed the previous year. Average ratings have been rounded to the ten's digit.

declines. 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 (this practice is usually called "sandbagging"). Nonetheless, the use of the rating floor is at odds with the principle that ratings are predictors 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, which is shown in Table 2. In January 1997, the mean rating among established players was about 1390, and in January 1998 the mean was 1340, a drop of 50 points. At the same time, if we examine the average ratings among established players active in both year, the average ratings are 1430 and 1440 for 1997 and 1998, respectively, resulting in an increase in 10 points.

How can we make sense out of the overall average rating among established players in January 1997 decreased from 1390 to 1340 in January 1998, and yet the average rating among players who were established in both years increases by 10 points? The answer lies in the flux of the established rating pool. At the end of 1996, 27592 players who were active over the previous year had established ratings. Slightly less than one-third of these players became inactive in 1997. These players had an average established rating of about 1310, as shown in Table 2. In contrast, there were 27563 players with established ratings in January 1998 who had been active during the previous year. Of these, about one-third were either inactive or had provisional ratings in January 1997 (corresponding to the second and third rows of Table 2). The average established rating for this group in January 1998 was 1130. In addition to retaining 18498 players from January 1997 to January 1998 who experienced a 10 average rating increase, the established rating pool lost a group of players with an average rating of 1310, and gained a group of players with an average rating of 1130. The net effect of this movement of players into and out of the established rating pool was an average rating decrease of about 50 points.

Towards an Improved Rating System

The USCF Ratings Committee has recognized many of the difficulties described here, and pending changes in the rating system are reported in Jones and Glickman (1998). One of the most significant is a revision of the treatment of unrated and provisionally rated players. The existing system treats these by methods with no theoretical connection to the underlying model or the rest of the system. The new system will impute initial ratings based on the best available information, which may be a rating in another system, or the overall rating-age relationship. Players with a small number of games (fewer than 8) will have ratings computed by an iterative algorithm which finds the rating which matches expected score to observed score.

The choice of the multiplier K in the current established rating formula is based purely on the player's rating; higher ratings go with smaller K, reflecting the greater stability of ratings for stronger players with established ratings. In the new system, it will be a function of the number of games played in addition to the player's current rating. Players with more than 8 games will have a measure of uncertainty of their rating. The multiplier K in the update formula will be a function of this measure (as well as the number of games in the event). The greater the uncertainty indicated by this measure, the larger the value of K.

With the greater availability of tournament data and better computational resources, examining and testing alternative models for chess strength is much more feasible than in the days when Elo was developing his system. It may be realistic after all to expect a chess rating system that truly predicts performances.

References

Bradley R. A., and Terry, M. E.(1952). “The Rank Analysis of Incomplete Block Designs. 1. The Method of Paired Comparisons.” *Biometrika*, **39**, 324–45.

Elo, A. E.(1978). *The rating of chess players past and present*, New York: Arco Publishing.

Glickman, M.(1998). “Parameter estimation in dynamic paired comparison experiments,” To appear in *Applied Statistics.*

Good, I.J.(1955). “On the Marking of Chess Players,” *Mathematical Gazette*, 39, 292-296.

Guare, J.(1990). *Six degrees of separation*, Random House, New York.

Jones, A. and Glickman, M.(1998). “The United States Chess Federation Rating System: Current Issues and Recent Developments,” to appear in the *Proceedings of the 1998 Joint Statistical Meetings: Sports Section.*

Zermelo, E.(1929). “Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung,” *Math. Zeit.*, 29, 436–460.

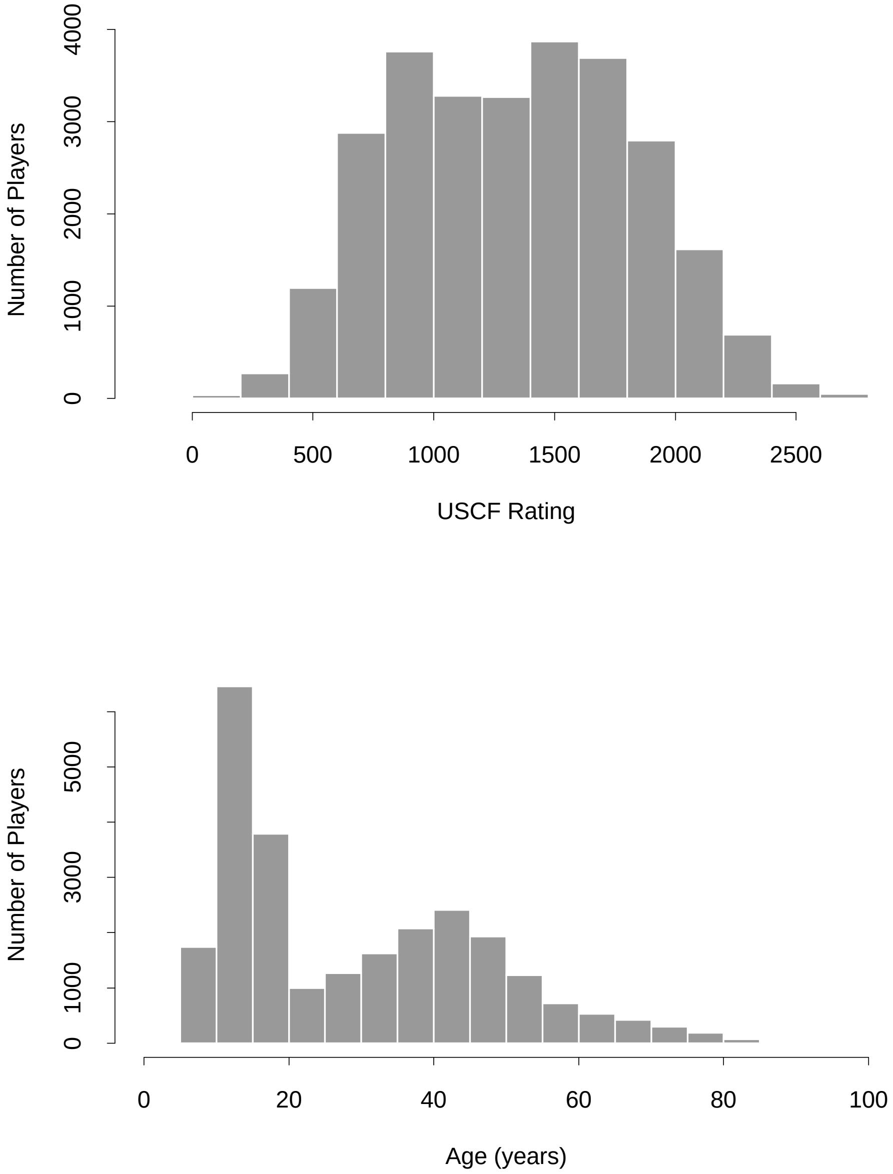

Figure 1: *Top:* Distribution of Established USCF ratings, January 1998. The histogram represents 27563 players who were active competitors (played at least one tournament game) in 1997, and who have played at least 20 tournament games in their careers. *Bottom:* Distribution of ages of established tournament players, January 1998.

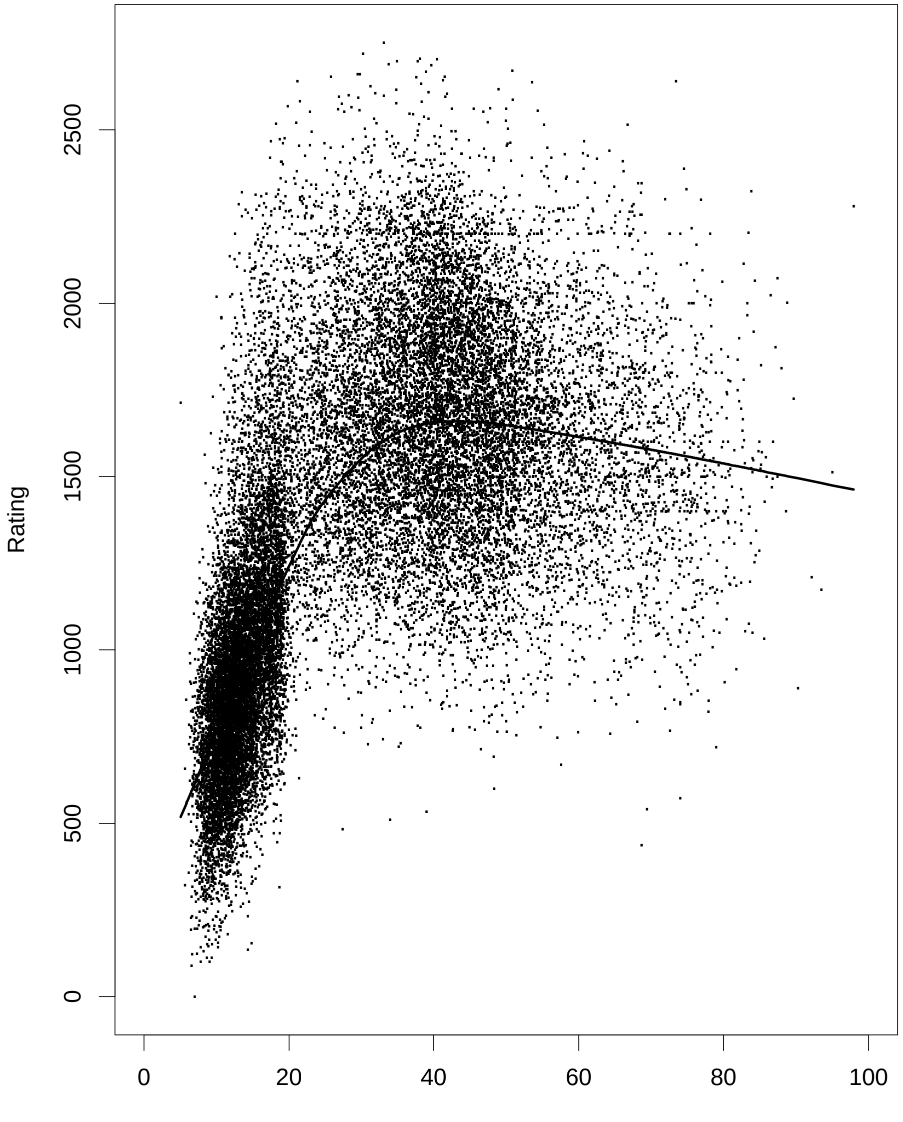

Figure 2: Scatter plot of USCF established ratings against players' ages in January 1998. A locally-weighted scatter plot smoother (lowess) is superimposed on the plot.

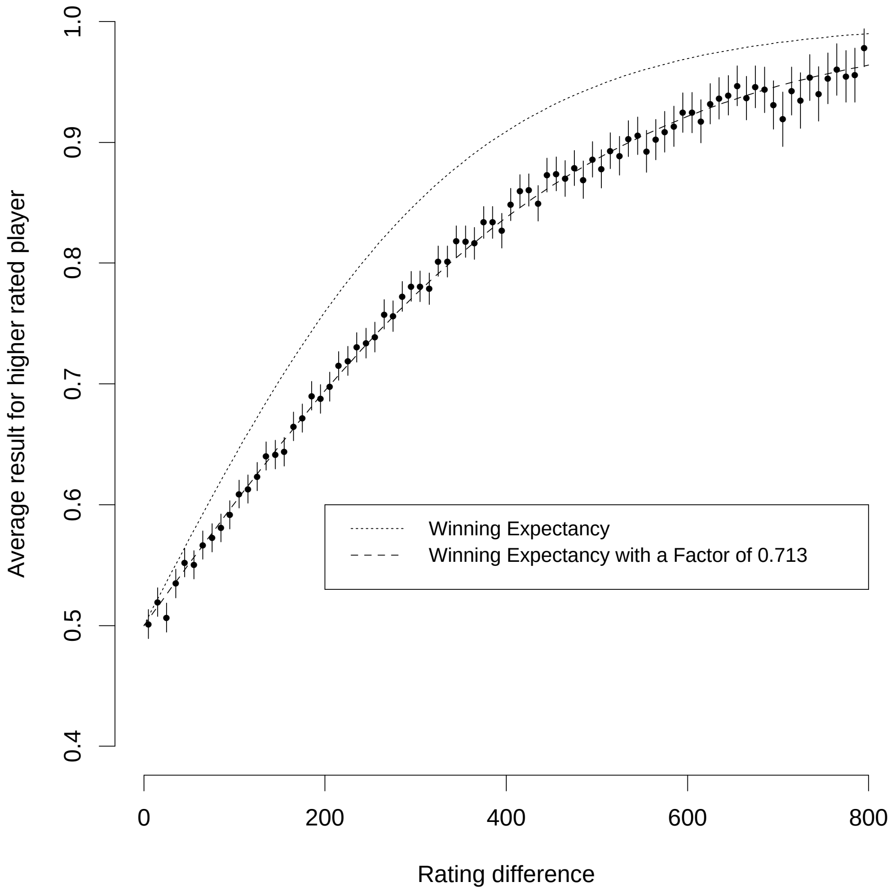

Figure 3: Summary of 225,621 rated USCF tournament games. Both players in a game must have had established ratings in 1997 to be included in the sample. The sample is partitioned into groups of players according to their rating difference (0-10, 10-20, ..., 790-800). For each rating difference group, the dot represents the average score of games relative to the higher rated player. The vertical bars represent 95% confidence intervals. The values on the dotted line are the expected scores calculated from the USCF winning expectancy formula. The values on the dashed line (which trace through the segments) represent a better fitting model.