ESSAY

Chess Rating Systems

Mark E. Glickman

The creation 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 assumed that a rating of 2000 would be equivalent to scoring 50% in a U.S. Open Championship. Elo's system, however, added considerable statistical sophistication.

The International Chess Federation (FIDE) adopted Elo's rating system in 1970. Since that time, the system has been adopted with various modifications by many national chess federations. Today it is hard to imagine tournament chess without a rating system.

Why Rate Chessplayers?

Chess rating systems have many practical uses. For pairing purposes in open tournaments, a tournament director wants to have some idea which players are considered the most likely candidates to win the tournament so he can try to avoid pairing them against each other in the earlier rounds of the tournament. Ratings are also used for tournament sectioning and prize eligibility. In most U.S. Swiss-system tournaments, only players of specified rating ranges can compete for section prizes.

Mark E. Glickman is a USCF master and chairman of the USCF Ratings Committee. He received his Ph.D. in Statistics from Harvard University. He is Assistant Professor of Mathematics at Boston University, and lives in Cambridge, MA.
The author thanks Chris Avery, Andrew Metrick, Ken Sloan, and Alan Losoff for their assistance in the preparation of this article.

Ratings can also be used as a qualifying system for elite tournaments or events. Invitations 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 importance of using ratings for such purposes can best be understood by considering the chaotic situation before ratings existed. In the days before ratings, it was not possible to view chessplayers' strength objectively, and invitations to important tournaments were typically based on players' reputations. When the young José Capablanca was invited to play at San Sebastian 1911, established masters like Ossip Bernstein and Aron Nimzovitch derided him as a "flashy amateur." Capablanca surprised both these critics by beating them and winning the tournament. Ironically, when Nimzovitch himself was invited to the great New York 1927 tournament, the Russian player Efim Bogulyubov said, "Everyone knows that he is not a real grandmaster." Nimzovitch's plus score in the tournament belied his critic. At least Capablanca and Nimzovitch got the chance to vindicate themselves. In the bad old days before ratings, it was also easier for champions to avoid matches with their strongest rivals. It might have been harder for World Champion Emanuel Lasker to avoid a match with Akiba Rubinstein, and U.S. Champion Frank Marshall to avoid a match with practically everybody, if objective rating systems had been in place during the first third of this century. Modern rating systems provide objective measures of ability—though not perfect measures, as we shall see—that are accepted for most practical purposes by virtually everyone.

In the bad old days before ratings, it was easier for champions to avoid matches with their strongest rivals.

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. International players, too, must achieve minimum threshold ratings before FIDE will award the FIDE Master, International Master, and International Grandmaster titles.

One of the greatest benefits of the rating system is that it allows competitors at all levels to monitor their own (and others') progress as they become better chessplayers. However—as will become clearer later—a paradox is involved in evaluating the movement of one's rating over time. This is because a rating only has meaning when compared against other ratings in the rating pool *at the same point in time*. Over time, the composition of the rating pool changes. As the *Oxford Companion to Chess* notes, the characteristic flux of the rating system "renders meaningless comparisons between players in different periods." Despite this evident fact, there has been much idle talk in the press and among chessplayers about Garry Kasparov "breaking Bobby Fischer's record," because Fischer's peak published Elo rating was 2785 and Kasparov—who is still active, of course—has been published as high as 2805. In fact, Fischer's and Kasparov's ratings

are only significant in relation to the ratings of their contemporaries. When Fischer peaked at 2785 on the July 1, 1972 FIDE rating list, Boris Spassky was a distant second on the list at 2660, 125 points back. As Kasparov himself has pointed out, no other player has so far surpassed his contemporaries since the inception of the FIDE rating list in 1970.

Types of Rating Systems

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 10 years, various forms of this system were used by different national chess administrations, including versions developed in the mid-1950s for the USCF by Kenneth 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 1950s 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 led the U.S. to adopt the Elo system in 1960.[1]

When Fischer peaked at 2785 on the rating list, Spassky was a distant second at 2660. No other player has so far surpassed his contemporaries.

The Elo 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 only on the outcomes of tournament games. When two players meet, the Elo system predicts that the one with the higher rating should win more often than the lower rated player. The bigger the difference in ratings, the greater the likelihood that the higher-rated player will win.

The entry "Elo rating" in *The Oxford Companion to Chess* notes, "The calculations behind a change of rating, and the proof of the calculation, are too technical to be included here." This article will discuss both the underlying ideas and the statistical formulae incorporated in the Elo system, including potential modifications.

While some other competitive sports organizations (the U.S. Table Tennis Association, for example) have adopted the Elo system to rate their players, non-probabilistic methods for measuring achievement remain in use. 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 com-

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

pared 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 Association of Tennis Professionals (ATP) ranking system awards "computer points" based mainly on the type of tournament (e.g., "Grand Slams," "Championship Series," etc.), total prize money in 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 a bridge rating, an ATP ranking can go down after repeated poor performances. The ATP system also incorporates the element of time, 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 points by not competing. This feature may be more appropriate for tennis than for chess or bridge, because one's tennis 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 include a time component, it does not guarantee smooth changes in rankings.

This article describes the basic principles of the Elo rating system, and how these principles are currently applied in various rating systems. The USCF rating system is the focus of attention, though much of the discussion extends to other implementations of the Elo rating system.

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

The Statistical Context of Chess Ratings

Statistical theory is a complex subject, but one that we will have to explore in order to discuss chess ratings. Readers with some statistical background will have an easier time following the discussion, but the main points should be clear enough to the layperson who reads attentively.

The problem of rating chessplayers falls into the area 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 seen as indicating a degree of preference through score differences; a game in which one team defeats another by a large margin conveys a greater degree of preference than a game in which the final score difference is close. Topics in experimental psychology such as choice behavior and sensory testing also involve paired comparison data. For example, the "Pepsi challenge" is a test to determine whether an individual prefers Pepsi-Cola to Coca-Cola.[3]

The problem of rating chessplayers falls into the area of "paired comparison" modeling in the field of statistics.

While Elo's name is by far the one most often associated with the development of the current chess rating system, the statistical theory underlying the system had been established well before his work in the late 1950s, and certainly before his well-known 1978 monograph.[4] The first work to give serious attention to modeling chess ability was by the mathematician Ernst Zermelo in 1929.[5] In this paper, Zermelo addressed the problem of estimating the strengths of chess players in an uncompleted round-robin tournament. Statistician Irving 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 imple-

3. A good overview of statistical modeling and analysis of paired comparison data can be found in Herbert David's *The Method of Paired Comparisons* (Oxford University Press, 1988).

4. *The Rating of Chessplayers, Past and Present* (Arco, 1978).

5. "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift* 29 (1929), 436–460.

6. "On the marking of chess players," *Mathematical Gazette* 39 (1955), 292–296.

7. "The rank analysis of incomplete block designs. 1. The method of paired comparisons," *Biometrika* 39 (1952), 324–45.

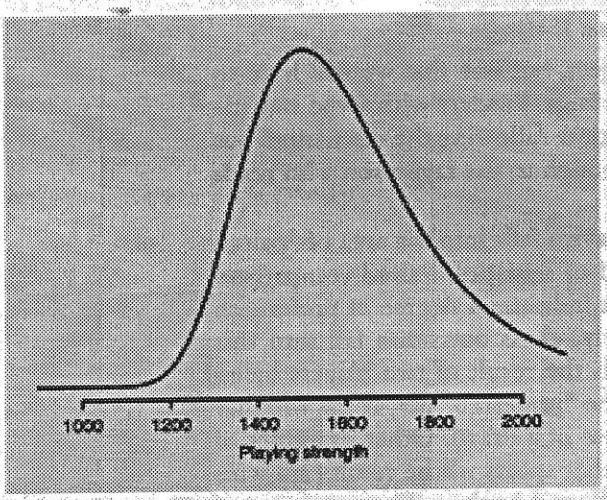

Figure 1 An extreme value distribution centered at a strength of 1500. Higher points on the curve indicate greater likelihood that a player will perform at that level.

mented versions of the 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." A statistician would then view a game of chess in the following way: 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 drawing 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 it does not guarantee his victory in every game. This is analogous to chess: The better player usually wins, but not always.

The Bradley-Terry model can be derived by making a particular assumption about the 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-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 and 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 probability of the first player having drawn a higher value from his or her strength distribution.

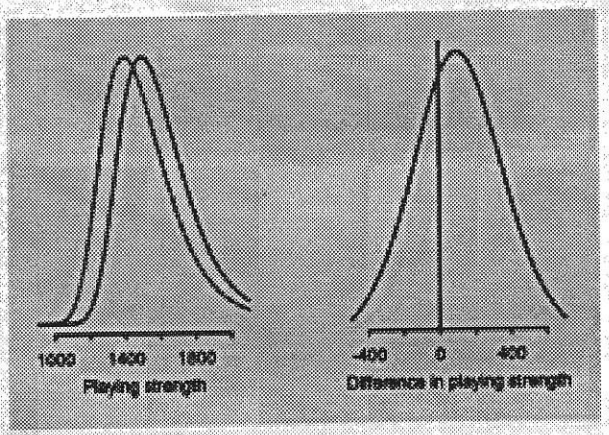

Figure 2 Left Two superimposed extreme value distributions, one centered at 1400 (dotted line) and one centered at 1500 (solid line). Right Logistic distribution of the difference between two players' individual performances. The area under this curve is the probability the stronger player will outperform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

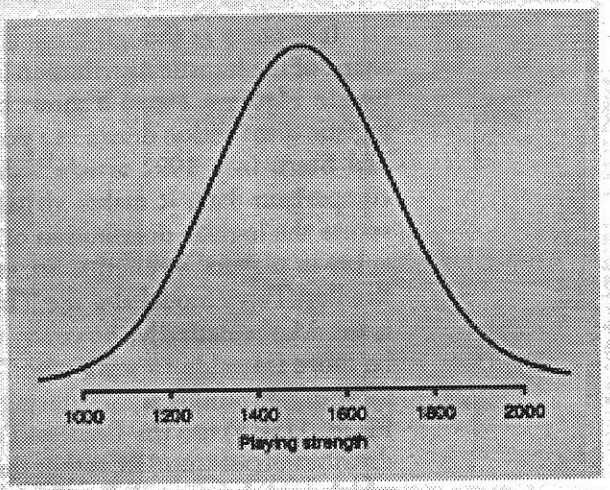

Figure 3 Normal distribution centered at 1500. As in Figure 1, higher points on the curve indicate greater likelihood that a player will perform at that level.

Even though the currently implemented system can be derived by assuming that a player's strength distribution is an extreme value distribution, Elo's chess rating system assumes that a player's strength distribution is a normal distribution (bell curve). **Figure 3** shows the curve for the normal distribution. The paired comparison model derived from the normal distribution is commonly known in the statistics literature as the Thurstone-Mosteller model, based on work by Louis Thurstone in the late 1920s,[8] and statistician Fred

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

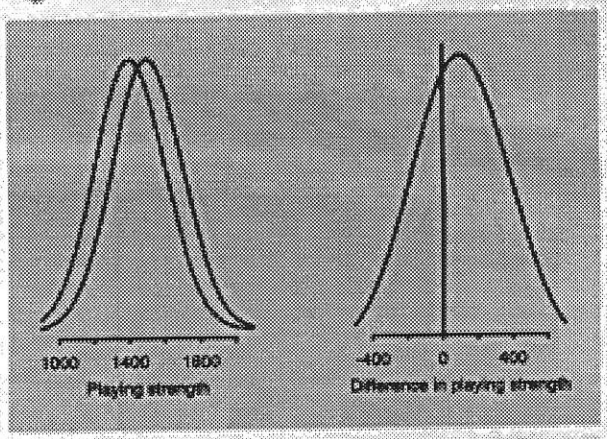

Figure 4 *Left* Two superimposed normal distributions, one centered at 1400 (dotted line) and one centered at 1500 (solid line). *Right* Normal distribution of the difference between two players' individual performances. The area under this curve is the probability that the stronger player will outperform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

Mosteller in the early 1950s.[9] In 1979 psychometricians William Batchelder and Neil Bershad, using the Thurstone-Mosteller model, extended Elo's model by formally modeling the probability of individual game outcomes.[10] One interesting feature of using the normal distribution to model a 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 dotted line. In fact, statistics professor Hal Stern in a 1992 article[11] showed 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.

Other models for rating chess performance have appeared in recent statistical literature. Statistics professor Harry Joe in a 1990 paper[12] examined the best chessplayers of all time with a model that

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

10. "The statistical analysis of a Thurstonian model for rating chess players," *Journal of Mathematical Psychology* 19 (1979), 39–60.

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

12. "Extended use of paired comparison models, with application to chess rankings. *Applied Statistics* 39 (1990), 85–93.

splits players' careers into "peak" periods and "off-peak" periods. This analysis was performed on a data set compiled by Raymond Keene and Nathan Divinsky.[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, Joe wrote an article in 1991 that derived axiomatically a general framework for a rating system, and showed that 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.

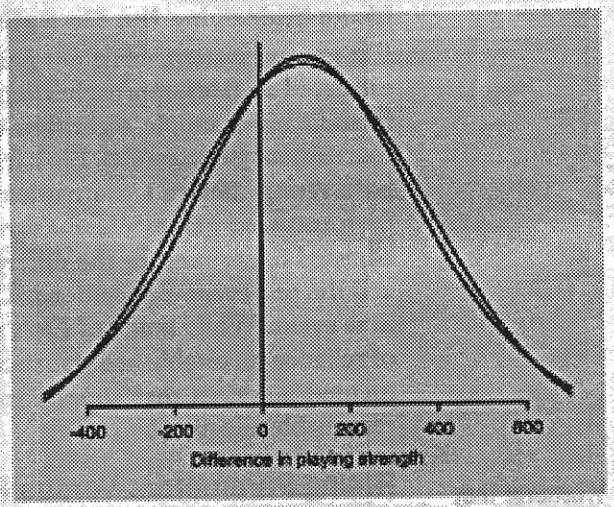

Figure 5 Two superimposed distributions of the difference between two players' performances—the logistic distribution (solid line) and the normal distribution (dotted line). For practical purposes, the two curves are indistinguishable.

Paired comparison theory has most typically been devoted to problems of modeling judges' preferences among a set of objects. While the game of chess, and most other games involving two 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 of the problem. My own Ph.D. thesis (Harvard University, 1993) developed 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, I analyzed the results from the World Cup tournaments of 1988–1989 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.

13. A prototype of this data set appeared in Keene and Divinsky's *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. "Dynamic paired-comparison scaling," *Journal of Mathematical Psychology* 36 (1992), 185–212.

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

Ideas Underlying the Elo Rating System

Elo's rating system, while not going to the same level of mathematical detail as later approaches, makes an important contribution by introducing a simple algorithm to adjust players' ratings based on tournament game results. Elo's framework is quite appealing: players have ratings before a tournament which, in principle, 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 system 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.

Rating Parameters Versus Rating Estimates

Whenever they analyze data, statisticians make a clear distinction between "parameters" and "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 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 "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 (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.

The distinction between estimates and parameters is rarely, if ever, made in the context of chess ratings.

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 num-

ber 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 game 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 (although 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 player can be expected to perform worse than his published rating would lead one to believe.

A computed chess rating is really an estimate of the player's rating parameter, that is, the player's average strength.

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

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

where the score of a game is 1 if player A wins, ½ 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

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

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 some of these games will be draws.

The paradox, of course, is that this formula applies only to rating parameters, which we can never 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.

Analysis demonstrates that the expected score function computed on estimated ratings does not describe the data.

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 8,300 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 Figure 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 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 fact, if a rating parameter is estimated

19. The 95% margin of error is an estimate of the error in using the sample average to approximate the true (population) average. In particular, 95% of new samples would have the true average within the given range. Shorter segments indicate, to some extent, larger samples.

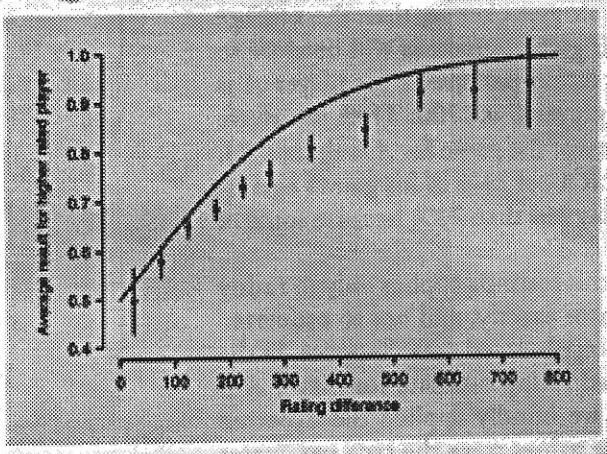

Figure 6 Summary of 8329 rated USCF tournament games. Both players must have competed in at least 20 tournament games to be included in the sample. The sample is partitioned into groups of players according to their rating difference (0–50, 50–100, 100–150, 150–200, 200–250, 250–300, 300–400, 400–500, 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 calcuated from Elo's expected score formula.

with error from player to player, we *should* expect the expected score formula to overestimate the observed outcomes. This is actually a statistical property of the expected-score formula. To understand this point, suppose 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 fact 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 practice, 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 than 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 statistical phenomenon is likely to be the main explanation for the behavior in Figure 6.

Updating Ratings

Because it is impossible to *know* a person's rating parameter *exactly*, the only hope is to *estimate* the parameter *accurately*. 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, it 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

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

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 can be understood as follows. 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, $(S - S_{exp})$. 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 was expected to score 3 points out of a five-round tournament given the opponents' pre-tournament ratings but 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 respective 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 four-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]

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

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 $r_{post} = 0.947r_{pre} + 0.053r_{perf}$. 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 formula (2) and tracking the position of a moving 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 formula (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 for estimating 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 before entering a tournament.

The USCF and FIDE have implemented systems to compute initial ratings using different sets of formulas. The resulting estimated ratings are often called "provisional ratings." As the name implies, we do not place great confidence in provisional ratings be-

cause 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 during which a player's rating is considered provisional. In the current implementation of the USCF rating system, this is a problem. Because no limit is put on the time one's rating remains 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 can be a problem when 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 used 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. When this player competes in a tournament, formula (2) is applied using the prior rating as r_{prev} and the attenuation factor K is set to be very large (e.g., 150) to give substantial weight to the performance. For a four-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 on 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 20 chess players, all possessing the same average strength, competes in a single-round-robin tournament, and the winner achieves a score of 14 points out of 19. Suppose also that the

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

player with the worst results obtains 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. If these 20 players were to compete in a second single-round-robin tournament, it is likely that the results of the winner from the first tournament would 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 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 overperformed 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 the performance of an average player. 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 be applied directly 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

This section will discuss the implementation of some of the leading chess rating systems currently in use, including the USCF, FIDE, and PCA scales.

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, tailored 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 in-

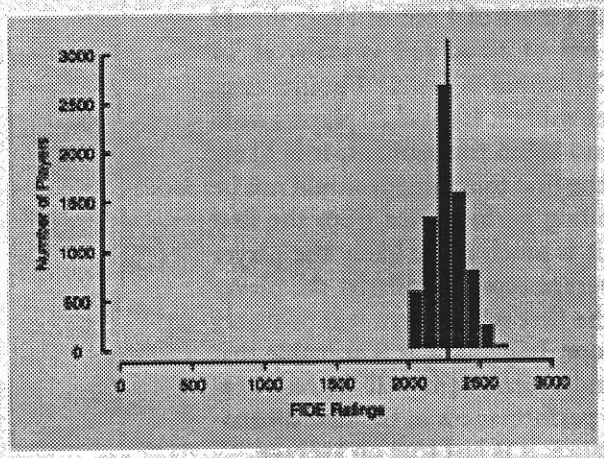

Figure 7 Distribution of FIDE ratings, July 1994. Players who competed in at least one FIDE-rated game in the previous six months are included in the sample.

corporate slightly different updating algorithms, it is not surprising that a FIDE rating will not correspond exactly in meaning to a USCF rating. As will be discussed later in this article, USCF ratings went through a period of deflation in the 1970s. Accordingly, corrective measures were adopted by the USCF. As of this writing, USCF ratings are somewhat higher than corresponding FIDE ratings. That is, a currently active player with established USCF and FIDE ratings will probably be rated somewhat higher on the USCF scale.

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 Elo's original 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] 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 substitute for Elo's original formulas.[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

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, Past and Present*.

24. For example, if a player rated 2005 competed against opponents rated 2600, 2600, 2600, and 2005, he would be expected to score about 15%, whereas the FIDE formula would yield an expected score of about 7%.

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.

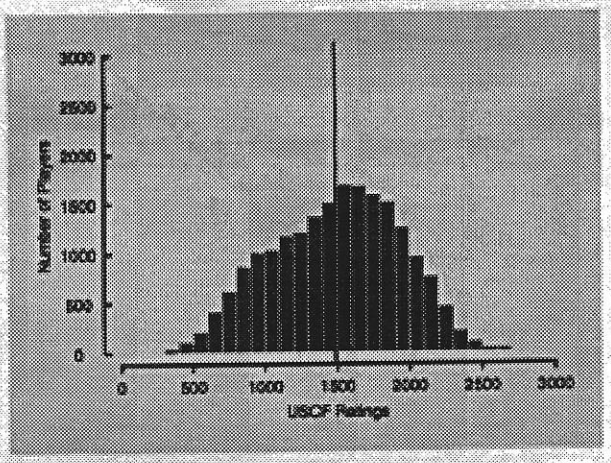

Figure 8 Distribution of USCF established ratings, July 1994. Players who competed in at least one USCF-rated game in the previous six months are included in the sample.

The USCF rating system, which assigns ratings to all competitors in USCF-governed tournaments, does not require 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 was 1490. USCF established ratings ranged from a low of 45 to a high of 2763. About 96% of all USCF established players had ratings less than 2200, as compared to FIDE's 23%.

A common misconception about the rating system is that players' ratings follow 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 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 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. There are 484 players with ratings on both the July 1994 FIDE and USCF rating lists. Among these 484 players, only players that had established USCF ratings and had played at least 6 FIDE-rated games in the prior six months before the publication of

25. The lowest rating a player can earn in the current USCF rating system is 0.

26. For example, the article "Ratings—Some questions answered" by Gerry Dullea in the December 1979 issue of *Chess Life & Review* made such a mistake.

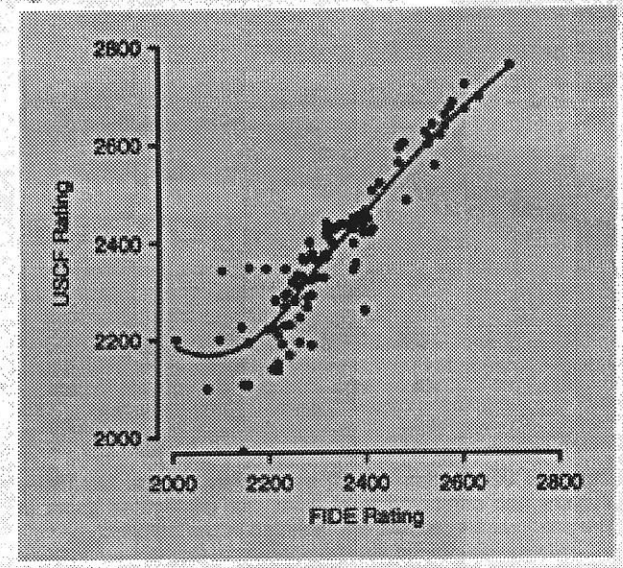

Figure 9 Plot of USCF ratings against FIDE ratings for 211 players common to both July 1994 rating lists. Players who played at least six FIDE-rated games in the previous six months and at least one USCF-rated game in the previous six months, and had achieved an established USCF rating, were included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

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

It turned out that most players had higher USCF ratings than FIDE ratings. **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 smoothing" that ignored unusual points (e.g., the player with a 2300 FIDE rating and an 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 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-2500s, and then declines once more to a difference of 65 to 70 in the high-2600s. 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 five games in six months) to appear in the analysis. Among the 273 players with USCF ratings that played that often, the USCF ratings tended to be much higher compared to the corresponding players who played fewer 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 system that calculates their "Intel World Chess 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.

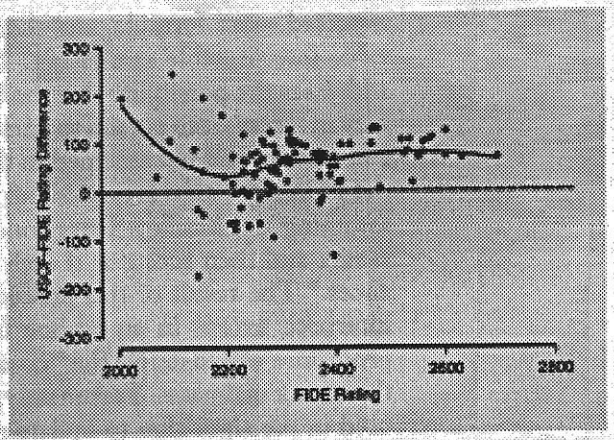

Figure 10 Plot of USCF–FIDE rating difference against FIDE ratings for 211 players common to both July 1994 rating lists. The criteria for inclusion in the sample are the same as those for Figure 9. The scatterplot smoother demonstrates that the USCF–FIDE average rating difference depends on a player's 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.

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 that 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 "variance" computation involves calculating the average squared deviation of each game result (1, ½, 0) from its expected game result using the

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.

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 the algorithm to update ratings, so the "variance" computation is not relevant to the predictive ability of PCA ratings.

The PCA system downweights games linearly. This may be an area for improvement.

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 systems 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 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 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

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

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

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 behind the Elo updating scheme is this: Rather than save all past game results and compute a rating based on all the data each time a tournament is completed, extract only the pre-tournament summary information and combine it with information from the tournament 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 re-compute ratings from scratch.

Rating System Characteristics

This section will discuss various factors that can affect the accuracy and reliability of ratings, including time controls, regional variation, and the passage of time.

Varying Time Controls

One of the newer features of the USCF rating system stems from the formal introduction of "quick chess," which refers to games where the time control for a game is shorter than 30 minutes per person for the entire game. In the late 1980s, it was debated 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.

People who perform better at quick chess may be demonstrating a different ability.

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 that 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 quicker at calculating tactics 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 would combining the rating systems for over-the-board and correspondence chess.

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. Nor is it 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 they can stand 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 2½ hours. When a player competes in a tournament with a time control in between these two rates of play, *both* ratings would be 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 2½ hours. Under such a system, a player might approximate his or her rating at various time controls by taking appropriate weighted averages of the two ratings. Of course, this system would require a further conjecture about the weights attached to the two ratings, so implementing such a system might be difficult in practice.

The rating system is only accurate on a "within-region" level.

Regional Variation in Ratings

The title of the recent play by John Guare, *Six Degrees of Separation*, 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.

The notion of being able to trace paths that connect players has direct relevance to measuring chess ability. No claim is made here that any two players have competed via six degrees of separation, but it can be asserted 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 local weekend tournaments attract roughly the same players, so that these same 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 likely never to have competed, rarely to have played opponents in common, or even to have played 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, assume two groups of tournament players. The members of each group only compete among themselves, and each group has an average rating of 1500. Also suppose that the players in the first group improve faster than those in the second group. After a period of time, both groups will still have an average rating of 1500, but a player rated 1500 in the first group will likely be notably better than a player of the same rating in the second group. However, if the players in each group only compete among themselves, then we cannot possibly determine that the players in the first group are better players on average than those in the second group *through their ratings alone*. Some connection is needed between the two groups in order to recognize a difference in abilities.

Ratings of scholastic players may be poor predictors of their performances against adults.

A situation like the foregoing, in which members of a group compete only among themselves, occurs frequently in scholastic chess. At the beginning of their chess careers, scholastic players may happen 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. 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 local situations are not as extreme as 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 claims by certain regions that they are systematically underrated relative to players in other regions.

The only 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 players finish the tournament, they bring back to their communities slight adjustments in their ratings that reflect the overall strengths of their opponents in other communities. Similar adjustments occurs when players move

from one region 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 solve the problem completely.

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. As they gain 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 means improvement?

Relating increases or decreases in one's rating over time to change in ability is a 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" examines 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 with above-average ratings. The net effect of this flux of players is a decrease in the overall average rating.

Elo argued that the average rating among rated players has a general tendency to decrease over time.

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 explanation of rating deflation can be tightened. Specifically, the existence of rating deflation requires two features of the rating system. The first is that players' abilities, on average, improve over time. We should not take for granted 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 due to competition against the underrated players.

In the mid-1970s, 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 an increasing discrepancy between 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-1980s, 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.

The updating formulas for the USCF rating system have been modified to combat rating deflation.

In the late 1980s, 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.[30] If, for example, a player's highest attained rating was 1871, then the player's rating could not decrease below 1600. More recently, the rating floor has been raised 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 rating floors argue that they will not only combat the natural tendency of rating deflation, but will actually encourage chess tournament participation because they prevent 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 against weaker players and win large cash prizes.[31] Nonethe-

30. The highest attained rating for every player only began to be recorded after the inception of the rating floors.

31. This practice is usually called "sandbagging."

less, 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 whenever a player at his rating floor loses a game (or draws a game against a lower-rated opponent). It is also possible that players at their rating floors may have misplaced incentives since they have nothing to lose: that is, some of them may adjust their styles by purposely playing more recklessly in the hope of winning with less effort, especially against higher-rated opponents. 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 that include players who had tournament games rated over the past year. In the January 1993 list, the mean rating of 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. The table below summarizes mean USCF ratings broken down according to players' statuses in 1993 and 1994.

Status 1/93	Status 1/94	Rating 1/93	Rating 1/94	Rating Change	# of Players
Established	Established	1632.6	1641.7	+9.1	12233
	Inactive	1548.4	–	–	9670
Provisional	Established	1143.1	1184.4	+41.3	1910
	Inactive	1086.4	–	–	7933
	Provisional	1124.7	1138.3	+13.6	1772
Inactive	Established	–	1421.8	–	4393
	Provisional	–	990.4	–	10777

The first line of the table indicates that 12,233 players had established ratings in both January 1993 and January 1994. The average rating for these 12,233 players in January 1993 was 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. The table 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 estab-

lished 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, 21,903 players who were active during the year had established ratings. Slightly more than 44% of these players became inactive in 1993. These players had an average established rating of 1548.4. In contrast, 18,536 players who were active in 1993 had established ratings in January 1994. Of these, slightly more than 34% were either inactive or had provisional ratings in January 1993 (corresponding to the third and sixth rows). The average established rating for this group in January 1994 was 1349.9. In addition to maintaining 12,233 players from January 1993 to January 1994 who experienced a 9.1-point 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 both 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.[32] Possibly 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 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.[33] The magnitude of this effect is hard to estimate.

If the rating system were functioning properly, we would not expect a significant increase in established players' ratings from one year to the next. In particular, the 9.1-point 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.

32. Similar analyses were performed on data between 1988 and 1989, and between 1992 and 1993, and the same conclusions resulted.

33. In January 1994, approximately 8% of all active players with ratings between 1400 and 2200 were at their rating 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 final digits.

The argument that the provisionally rated opponents of established players are, on average, overrated is based on inference. It runs as follows. Clearly, provisional ratings are subject to great uncertainty, so that sometimes one would expect a provisional rating to overestimate a player's ability, and sometimes one would expect it to underestimate. If the provisional-rating system worked properly, the number of provisionally rated players whose average strengths were overestimated would equal the number whose average strengths were underestimated. If this were so, then among all contests involving a provisionally rated player and an established player, the average rating change among established players should be close to 0. 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. However, even when the provisional-rating system works properly, we would expect players' provisional ratings, in general, not to keep pace with their true average strength, but to *underestimate* it. This is because of the further assumption that provisionally rated players are generally improving at a more rapid pace than established players. If the provisionally rated players are, on average, underrated, then the established players should lose rating points overall. Obviously the reverse is happening, as the table on p. 88 demonstrates. We may infer, therefore, that provisionally rated players are not underrated but *overrated*. This inference provides evidence that the rating system may not be properly functioning.

We should not necessarily be concerned about changes over time in the average rating of tournament chessplayers.

Even though adjustments to the rating system have been implemented to counteract rating drift, it is worth pointing out that we should not necessarily be concerned about changes in the average rating of tournament chessplayers. It all depends on the goals of the rating system. The rating system by itself only makes assumptions about *differences* in players' ratings, not in their actual value. If 1000 were subtracted from (or added to) everyone's ratings, the rating system would still be just as valid, because differences in players' ratings would remain the same.

That being said, it is obvious that 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, popular opinion believes that "1800 strength" this year should connote the same ability next year, five years from now, and 20 years from now—and if somehow this does not happen, then something is wrong with the rating system. Unfortunately, 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 observation has been made by writer and computer consultant John Beasley,[34] 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 these factors prevent measuring absolute changes in ability from game outcomes. Suppose, for example, two players, both 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, by intense study and informal practice. However, their ratings are both still 1500, because they have not played any rated chess games. They compete again, and again each scores 5 out of 10. Even though both players have improved vastly, we cannot detect this, because their ratings will each remain unchanged at 1500.

Ratings can be used only to describe relative abilities, not absolute abilities.

Although it may not be possible to guarantee that a given rating will mean the same thing over time, it is possible to set a goal of maintaining certain characteristics of the overall rating pool. One possible goal might be to force the median rating to a specified level, or some percentile of all active players to a specified rating by periodically adding a fixed amount to all ratings. Suppose, for example, that a median of 1500 is desired. Then 50% of all players will have ratings above 1500 from year to year. This would allow a player to compare his rating with the 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 adding 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 to meet this condition. 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 would mean that a player's rating might change due to an overall pool adjustment even when he or she is not competing.

Another idea that has been proposed is to align one rating scale to match another rating scale that is considered more universally acceptable. For example, the USCF has often considered aligning its rating scale with the FIDE scale, by updating USCF ratings periodically so the two scales have the same absolute interpretation. However, neither the FIDE system nor any other system in existence guarantees stability in its rating scale or its rating system. With the decision in 1993 to exclude Gary Kasparov and Nigel Short from the FIDE rating lists, FIDE opened itself to charges that its rating sys-

34. *The Mathematics of Games* (Oxford University Press, 1989), 60.

tem was vulnerable to political manipulation, which alone would seem to disqualify it from being a "gold standard" of rating systems. A further argument against aligning two rating scales, such as the USCF and FIDE scales, is that the link from one scale to the other might be based on a small number of players, so the alignment might fluctuate primarily due to the imprecision of the estimated conversion between the two scales. Also, in trying to gain control over the USCF rating system, it is unappealing in principle to impose a condition on it that depends on information from another system over which the first system has no control.

Finally, one possible direction of effort is to develop tools, based on factors external to the rating system, to make ratings connote the same ability over time. One basic idea borrows from "item response theory" in educational testing. The Scholastic Aptitude Test (SAT) taken by many high school juniors and seniors has been constructed so that current students' performances can be compared to students' performances of the past. The Educational Testing Service does this by including a number of test items common to different exams. Thus individual exams are "linked" together by common test questions. Through these links, paths can be inferred that connect students of the past to students of the present via statistical models. Any given SAT score thus connotes the same ability today as in the past.[35]

FIDE has opened itself to charges that its rating system is vulnerable to political manipulation.

This approach can be applied to rating chessplayers in several different ways, though the merit of any of these methods is certainly arguable. One idea is to make use of chessplaying software. Because the chessplaying 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 could first be accurately estimated by having them compete against each other, as well as having them compete against a wide selection of humans. These ratings could then be used as fixed "anchors" in the rating system. Periodically, these chess programs should be entered into tournaments. The results of competition would determine the magnitude of any overall ratings drift. The drift could 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 under 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

35. As of 1994, the SAT was no longer designed to connect scores to the past in this manner. Instead, it now determines scores that correspond to percentiles of the current population taking the exam.

against chess programs than they do against other humans. A performance against chess programs may not translate to an equivalent performance against humans of the same ratings. Also, implementing such a procedure of having computers play against humans regularly might be impractical and expensive.

A variation of this theme would consist of periodically identifying groups of players who seem to demonstrate stable abilities, and using them as anchors in the rating system for a certain length of time. It would be essential to prevent people from knowing which players were being used as anchors. Candidates for anchors would be those players who compete regularly without significant rating fluctuation. Such players might be used as anchors for 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 is that it is difficult, if not impossible, to assess *a priori* that a player's ability has reached equilibrium. This difficulty is exacerbated by the well-known phenomenon of "plateauing," in which a player's ability—and therefore his rating—may stay the same for months or years, and then jump up dramatically as a result of intangible factors such as additional study time, more experience, more confidence, or a change in openings or playing style.

Finally, a more rational approach to creating a system in which ratings connote the same ability over time involves designing a chess test to measure chess ability, and then designing a statistical model to predict chess ratings from the test. A series of chess questions could be constructed to test ability in all phases of the game. A sample of rated chessplayers would take 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 to a different sample of players to see how the ratings derived from the test results differed from the actual tournament ratings. Based on these differences, an adjustment could be applied to all ratings to preserve the constancy of ratings over time. This approach, while making use of a source other than game results to measure chess ability, has the fringe benefit of identifying the aspects of chess that separate weak chessplayers from strong ones. On the downside, assessing the accuracy of the test becomes a new source of variability, and could increase the difficulty of measuring playing strength. In any case, designing and administering such a test and performing statistical analysis of the data could be expensive to carry out correctly, and for that reason among others might not be in the interests of chess organizations.

Improving the Rating System

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

Advantage Due to Color

The Elo rating system as currently implemented appears to function reasonably well.

It is commonly understood that having the white pieces confers an advantage. Elo estimates that White has a 1.33 times better chance of winning than Black.[36] In my Ph.D. thesis, I used results of the World Cup tournaments of 1988–89 to estimate that, among top masters of similar abilities, White has a 1.56 times better chance of winning than Black. 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 of having the white pieces 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 the board as White, the value of 80 is automatically added to every value in his box. This is a straightforward mechanism to describe how a statistician might model the advantage to having the white pieces.

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

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

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-

36. See Section 8.93 of *The Ratings of Chessplayers, Past and Present*.

rating-point advantage for White.[37] This formula has strong connections to a model postulated by statisticians Roger Davidson and Robert Beaver in 1977.[38] 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.

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 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 to achieve with the white pieces than with the black.

With such a large advantage to White, it makes sense to consider incorporating color information.

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 adding 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.[39] It would be more accurate to say that the information regarding draw probabilities is just as available as information regarding winning and losing, 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.[40] 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 in a particular game exhibit roughly comparable playing strengths, then the outcome of the game should be a draw. Rao and Kupper describe the procedure for esti-

37. The PCA determined this value by finding the average score for White in a database of over 100,000 games. Their analysis, however, did not take the players' strengths into account, so it is likely that the true advantage of playing White is less than 32 points.

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

39. See Section 8.91 of *The Rating of Chessplayers, Past and Present*.

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

mating 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

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

The probability player B defeats player A can be computed by substituting R_B for R_A in the above formula. 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 ratings is the same. Davidson and Beaver, besides describing how to incorporate the advantage of playing White into the Bradley-Terry model, also describe how to extend Rao and Kupper's model for drawn games to incorporate the advantage of 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 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 currently used.

Even though the system now in place only calculates the expected outcome of a game, and is not directly connected to a simple probabilistic mechanism like randomly 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. Although potentially valuable information is lost by not modeling individual-game probabilities, there is a realistic chance that the model does not accurately describe frequency of game outcomes anyway.

Incorporating the Uncertainty of Ratings

Some players' ratings are more poorly estimated than others. This inevitable feature of the current rating system has mostly been ignored, except in specific instances.

The problem can arise 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. Second, players who have not competed in tournaments for an extended time may have become either weaker or stronger, so that their ratings are less reflective of their true average strength. The rating system currently makes no distinction between established players who compete regularly and those who compete sporadically. In both cases, changes in the procedure for updating ratings would be required to incorporate the uncertainty in estimating ratings.

Uncertainty also occurs when an organizer is late in submitting a tournament report to the chess federation office. The USCF rates events in the order it receives reports, without regard to the actual dates of event. Suppose two events, *G* and *H*, occur separated by two months with *G* occurring first. If the organizer of event *H* submits a tournament report promptly, but the organizer of *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 of 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.

Some players' ratings are more poorly estimated than others.

The problems stated above can be alleviated in several ways. One approach allows *K* in the updating formula to be a function of time since the player 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 exceptions to this rule occur when a player's rating is from 2100 to 2399 (when *K* becomes 24), and when a player's rating is 2400 or higher (when *K* becomes 16). While the origin of this modification to the Elo system is not well-documented, one reason for its adoption is that players with high ratings are hypothesized to have abilities that do not change much over time. Therefore *K* should be lower to reflect this stability. However, the argument to base *K* on rating is not compelling.

When *K* is large, past performances are effectively downweighted relative to the current performance. Two cases when it might be useful to have a larger-than-usual *K* are when a player has a rating

based on very few games—so that 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. It might be appropriate to have a lower value of K when a player is competing regularly, because his or her ability is likely to be well-represented by the player's pre-tournament rating. Also, perhaps K should be low when an organizer has submitted a tournament report much later than the tournament's ending date if more recent performances have been rated. For example, if a tournament was completed in June 1992, but the results were not submitted until August 1993, these results should be given relatively little weight in comparison to results from a much more recent event. When the players' ratings are updated, little weight should be given to this performance from a year earlier.

When changing K in the updating formula to account for the uncertainty in a player's pre-event rating, a similar modification is necessary for updating the 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 than his rating would indicate, but with a poorly estimated rating, in which case the established player should not lose as many rating points. We conclude that a player who competes against an opponent whose K is large should gain or lose only a fraction of the usual number of points.

A 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.[41] 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, suggesting 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 does, this

41. This method has been adopted by the developers of the Free Internet Chess Server (FICS) for its rating system.

extension can give a plausible interval of values of the rating parameter, with the interval being wider for players whose rating estimates are more uncertain.

The differing measures of variability from player to player have consequences for 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 relatively 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 other player.

The passage of time has an effect on the variability of one's rating estimate. As more time passes, the measure of variability could be increased to reflect the extra uncertainty in one's ability. In fact, the system could 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 quickly. Furthermore, 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 precisely using integral calculus, but approximated numerically by a simple formula.[42]

No technical or theoretical principle demands that rating points be conserved.

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 equal the rating loss for the other. The size of the changes would depend on the variability of each player's rating. This might seem, at first, to violate some underlying principle that points in the rating system must be conserved, but this "principle" is a myth. No technical or theoretical principle demands that rating points be conserved. In fact, as argued earlier, it is a blind adherence to this principle that is partly responsible for rating deflation. Appropriately incorporating measures of variability into rating estimates is one way to tackle the problem of deflation.

Competing Incentives

One of the most important problems with the current rating system has little to do with its computational aspects or the validity of its

42. The details of the calculations are found in "An extension of the Elo rating system," an unpublished paper by the author.

assumptions. It has to do with players' perceptions of ratings and the consequences of those perceptions. While the implementation of a chess rating system has probably increased the popularity of tournament chess, it may also be responsible for driving some players away.

In the popular mind, 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 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 play for fear of losing more rating points. The view that declining ratings are a punishment or insult is a disincentive for players to compete. One could take an alternate view, that a lowering of one's rating merely indicates that a player was initially overrated, not that a player's ability is declining. However, the fact remains that rating changes often affect a player's pride or self esteem.

Chess rating systems have probably increased the popularity of tournament chess, but they may also be responsible for driving some players away.

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 from 1800 to 1999, the player is called a "Class A" player; if the rating falls from 2000 to 2199, the player is called an "expert"; if the rating falls from 2200 to 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 results. Similarly, when a rating drops below a class boundary, disappointment may result.

Even more consequential is that tournament organizers in the U.S. divide tournaments according to rating classes. Players whose ratings are just above a rating class boundary are prevented from participating in a lower class section, even though their ratings may be estimates of strength with high variability and their true strength might actually place them in the lower section. Dividing tournaments into sections by rating also creates an incentive for players to manipulate their ratings by artificially lowering them. They can accomplish this by purposely losing games in unimportant tournaments. The current design of 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.

In the last few years, the USCF has developed an additional system 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 by ratings. (Not many of these experiments have yet been carried out, however.) The title system does not intend to track players' abilities as the rating system is designed to do; instead it rewards players for incremental improvements in their performances.

The USCF title system is based on the principle that an exceptional tournament performance should be rewarded, but a poor tournament performance should simply be ignored. To earn an "1800 title," a player must achieve results in tournaments that exceed an 1800-player's expected performance by a certain margin. Under the current system, such a player would need to demonstrate five appropriately strong 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 result would have no effect on his four accumulated norms. The title system only rewards positive results and does not punish poor results.

The USCF has developed "titles" to function as a relatively independent reward system.

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 both 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 an Elo rating was not intended to be.

The USCF title system has strong connections to the system used by FIDE for awarding titles, such as the titles of grandmaster and international master. In the FIDE system, players must achieve outstanding results in events with 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 a higher rating. I believe that class designations of ratings should be stripped away and associated solely with titles to restore the unconfounded interpretation of ratings as measures of ability. The less attention players pay to their ratings, the less reason players will have to feel discouraged by rating decreases. Furthermore, titles provide players with an incentive to keep playing in tournaments without the risk of dropping down a class because they have lost rating points. As Macon Shibut, editor of *Virginia Chess*, has argued in an unpublished article "USCF Lifetime

Titles: A Good Idea, But Will It Fly?" the title system needs to become part of chess culture much in the same way that the current rating system has become.

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 built into 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 1960s, players' ratings were kept on index cards and updates were computed by hand. As membership grew and the number of tournaments increased, updating ratings by hand became a tedious task. Doing this today would be unthinkable. With more than 30,000 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-related functions. Fortunately, because ratings are now updated by 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 that players win chess games and how players' abilities change over time.

While we are thinking about how to make the Elo rating system more accurate and more useful, we should also consider putting it in its place as a tool for measurement and prediction. The title system should replace ratings as an incentive system and as a way of grouping players—at all levels, not just the international level. This would remove the pressure on the rating system to be a reward/punishment system, which it was never designed to be. ❧