

# Back to Basics in Chess Ratings

Royal C. Jones, Jr. ©2011

Second Edition: August 1, 2011

## Contents

1	Beyond the Elo System	2
2	Probability in Rating Systems	3
3	Ordinal Ratings	5
4	Interval Ratings	10
5	Ratio Ratings	12
6	Sequential vs. Simultaneous Ratings	13
7	Consistency	15
8	Established Ratings	20
9	Attenuation	22
10	Percentage Expectancy	25
11	A Revised Elo System	27
12	Cumulative Averaging and the Differential	28
13	The Berkin System	31
14	A Progressive System	33
15	Tests	34
16	Skeptical Conclusions	35
A	References	38
B	Downloads	39

# 1 Beyond the Elo System

The first attempt at a mathematical rating system for chess has been credited to the Correspondence Chess League of America in 1939. With a world war intervening, the influential Ingo System of West Germany followed in 1948, named by its originator Anton Hoesslinger (1895-1959) for his home town, Ingolstadt in Bavaria.<sup>1</sup> It establishes the basics of ratings with a simple formula,

$$R = ER_c - (Pct - 50), \tag{1}$$

where  $ER_c$  is the arithmetic average of the opposition ratings and  $Pct$  is the player's score in percentage points. (A peculiarity here, from the standpoint of subsequent systems, is that lower ratings represent greater playing strength.) The Ingo formula represents a notable advance in the rating of chess players, though it appears to have sprung primarily from Hoesslinger's intuition. The formal development of rating theory began about 1960 with the introduction of probability formulas. The originator of this idea was Arpad E. Elo (1903-1992), one of the founders of the United States Chess Federation (USCF), whose system was adopted in 1970 by the International Chess Federation (FIDE). The paradigm shift proved irresistible to mathematicians. About the same time that Elo was developing his system, similar ideas were afloat in Australia.<sup>2</sup>

The new formal theory of ratings was based on an implicit measurement of playing strength, necessarily implicit since there is no clear measurement of playing strength beyond the obvious facts of winning, losing, and drawing. The measurement would become explicit with the development of ratings, just as the notion of gravity had become explicit with Newton's formulas. As a professor of physics, Elo would no doubt have found this analogy appealing. It will be the burden of this treatise to show that ratings are measurements of playing strength in a figurative sense only. The simple fact is that ratings are *statistics*. The information they convey is based solely on the data provided by pairings and outcomes. To imagine that they represent some other dimension of playing strength, if only hypothetically, is to invite premature speculations about probability distributions, leading by a circular route to arguments for probability treatments based on the same distributions.

On the strength of probability theory Elo judged the Ingo and similar systems to be deficient because they were unwittingly based on a rectangular (uniform) distribution as a consequence of their linear formulas. The implication is that every rating system is based on a probability distribution and that the accuracy of a system is to be judged by the suitability of this distribution. Elo offered two complete systems, one based on the normal curve, another on the logistic. Apologists are quick to point out that there is little practical difference between the two systems, though the proposal of alternatives seems problematic by Elo's own standard. On the view that ratings are statistics we can hardly call any rating system invalid. We shall have occasion to call the Elo System cum-

---

<sup>1</sup>HW, "rating"

<sup>2</sup>E2, Part 1

bersome and not entirely coherent, but a judgment of invalidity would admit the mistaken standard it adopts.

By analogy with the commonly accepted scales of measurement, Elo distinguished three types of rating systems: *ordinal*, *interval*, and *ratio*. Since game scores in aggregate lend themselves to these scales of measurement, the classification is convenient for describing the various statistical methods that arise from rating theory and will be utilized in the following pages. But first the issue of probability will be revisited.

## 2 Probability in Rating Systems

The first assumption of the Elo System is that the chess performance of an individual player is a random variable that can be described by the normal curve.<sup>3</sup> But what is varying in this random variable?

As applied to a single game, performance is an abstraction which cannot be measured objectively. It consists of all the judgments, decisions, and action of the contestant in the course of the game. Perhaps a panel of experts in the art of the game could evaluate each move on some arbitrary scale and crudely express the total performance numerically, even as is done in boxing and gymnastics.<sup>4</sup>

Performance in this light is not a promising candidate for mathematical treatment, but there is a simple definition of performance that is as old as the game. If a player outperforms the opponent, the player scores a point; if the player is outperformed, the opponent scores the point; and if a draw occurs, the point is divided. This simple definition allows an interesting probability treatment. Under the heading “Sundry Theoretical Topics” Elo pointed out that the probability of a specific outcome in terms of wins  $W$ , losses  $L$ , and draws  $D$  can be calculated precisely as

$$P(W, L, D) = \frac{N!}{W! \cdot L! \cdot D!} \cdot P(\text{win})^W \cdot P(\text{loss})^L \cdot P(\text{draw})^D \quad (2)$$

if we know the probabilities  $P(\text{win})$ ,  $P(\text{loss})$ , and  $P(\text{draw})$ .<sup>5</sup> Let us consider the outcomes for ten games ( $N = 10$ ), where  $P(\text{win}) = .5$ ,  $P(\text{loss}) = .2$ , and  $P(\text{draw}) = .3$ , and let us express the results in terms of points scored. Since (2) must be calculated for 66 three-way partitions of 10, this is best done with a computer program (See Figure 1 below and “Downloads”).

Elo presented Formula (2) as an afterthought, but it provides the only convincing demonstration of the “first and basic assumption” of his system.<sup>6</sup> Contrary to the assumption of a constant variance, the variance here becomes increasingly narrow as  $N$  increases. As  $N$  goes to infinity, the distribution becomes

---

<sup>3</sup>E1, 1.31

<sup>4</sup>E1, 1.32

<sup>5</sup>E1, 8.9

<sup>6</sup>E2, Part 1, “Form Varies”

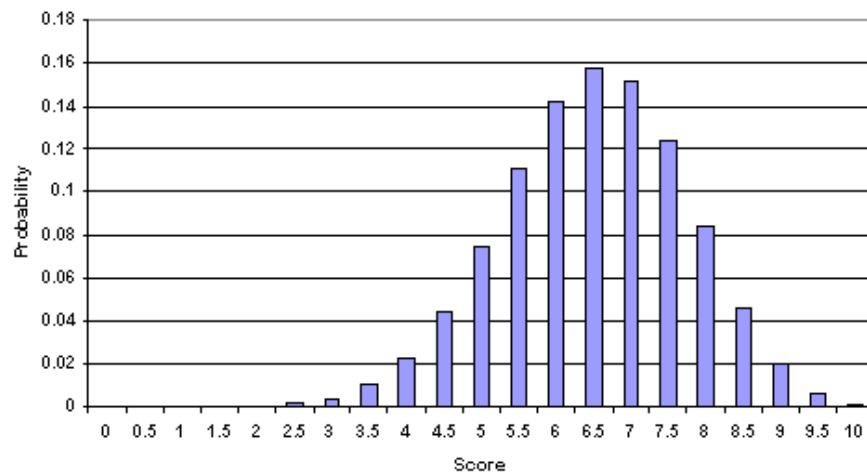


Figure 1: Score Probabilities for Ten Games

a vertical bar at the average score, 6.5. It is this long-term average that is the key to probability in rating systems. The problem then becomes one of relating the long-term average to rating difference.

Elo’s nebulous definition of performance leads to the central concept of his system: the Percentage Expectancy Curve, which is patterned on a well-known probability function, either the normal or the logistic. The Percentage Expectancy Curve relates percentage score to rating difference. Exactly how it does this is a crucial question. We are shown two overlapping distributions and are told that the shaded portion of one “represents the probability that the lower rated player will outperform the higher.”<sup>7</sup> Apparently, if a player’s performance is greater than that of the opponent, the player wins; if the performance is lower, the player loses. This argument, aside from the objection that it leaves draws completely out of account, makes distinctions in distributions that are already ill-defined. Outperforming the opponent seems equivalent in every respect to winning; yet this would suggest a binomial distribution, if not trinomial as in the illustration above.

It is sometimes noted in support of the Percentage Expectancy Curve that the distribution on which it is based appears quite often in natural phenomena, in everything indeed from IQ scores to errors of measurement. This fact was not lost on Elo:

Eminent mathematicians have tried many times to deduce the normal distribution curve from pure theory, with little notable success. “Everybody firmly believes it,” the great mathematician Henri

---

<sup>7</sup>E1, 8.23

Poincare remarked, “because mathematicians imagine that it is a fact of observation, and observers that it is a theorem of mathematics.” (Poincare 1892)<sup>8</sup>

This somewhat mysterious observation is explained by the prevalence of binary phenomena in nature. It applies, in any case, only to phenomena that are measurable and whose measurement is independent of the normal curve itself.

For Elo the Percentage Expectancy Curve was patently a probability function, and he would entertain no objection that it was not. Since a percentage score may be thought of as an estimate of probability, defined as a long-term percentage, there is reason enough to regard the Percentage Expectancy Curve as a function that relates probability to rating difference. In this broad sense, it is a probability function. More precisely the term applies to those functions that arise in probability theory from a mathematical analysis of variability, such as the normal curve or the logistic, and these we may call *true* probability functions. A function that merely maps probability to another variable without some justification based on variability analysis would thus be called an *arbitrary* probability function. Such functions include those based arbitrarily on a true probability function, which is the case of the Percentage Expectancy Curve. If the Percentage Expectancy Curve itself were a true probability function, it would be derived independently from distributions of rating difference, however these might arise, but these can hardly be known without a pre-established definition of ratings.

It need hardly be said that an arbitrary probability function may take virtually any form, including the linear form deprecated by Elo. Since the function is by definition arbitrary, it cannot be improved by mimicking true probability functions. Choosing the best statistic for a rating system will depend on criteria other than variability analysis. What follows is a belated attempt to put rating theory on a sound footing.

### 3 Ordinal Ratings

Ratings that are based on an ordinal scale typically take the form of ranks, though other forms, such as percentiles, are possible. The process of ranking is based on the dominance relation, indicated in a dominance graph by an arrow or *directed edge* drawn from the dominant to the subordinate player or *node*. Dominance in a competitive setting means prevailing over an opponent by some established criterion, such as a win or winning percentage. A competing field is represented by a complete dominance graph when all pairings are accounted for. In a partial dominance graph there are edges missing, corresponding to pairings for which there are no decisive results. The classic ranking problem is to assign ranks to the nodes of a dominance graph so as to minimize *violations*, in this case dominance by lower-ranking players. The problem has been shown

---

<sup>8</sup>E1, 1.39

to be NP-complete.<sup>9</sup> That is, barring an earth-shaking discovery, there is no ranking algorithm that is both exact and efficient for data of significant quantity. Rankings of a dozen or so players can be optimized by brute force, but as the number of ranks increases the degree of complexity increases exponentially. Finding a ranking that minimizes violations soon becomes impractical, even for a high-speed computer.

For some NP-complete problems, difficulties can be overcome by restating the conditions of the problem, though at the risk of trivializing it. It turns out that there is an efficient algorithm for ranking so as to minimize the algebraic sum of *directed rank differences*. A directed (signed) difference in rank is taken from the rank of the subordinate player to the rank of the dominant player as

$$\text{rank}(\text{dominant}) - \text{rank}(\text{subordinate}).$$

For expected results the direction is negative since the higher ranks are smaller. Such results are desirable from the standpoint of ranking because they tend to minimize directed rank differences. For upsets the direction is positive. The algorithm for minimizing the algebraic sum of directed differences in a ranking is simply to sort the players in descending order according to the difference

$$W_d - L_d$$

with respect to each player, where  $W_d$  is the number of opponents dominated by the player and  $L_d$  is the number of opponents who dominate the player. (The notation suggests wins and losses, but only as a mnemonic device. The subscript indicates dominance. The equivalent in a dominance graph is *outdegree* minus *indegree*.) The algorithm is applied recursively as a tiebreaker among players of equal rank.

It is difficult to prove the efficiency of sorting by  $W_d - L_d$  to minimize directed rank differences without being long-winded. A wordy sketch will suffice for the present purpose. Note that when a player moves up in rank, the sum of his directed rank differences is decremented for each player dominated by him (that is, the sum is decreased by  $W_d$ ) and is incremented for each player who dominates him (that is, the sum is increased by  $L_d$ ). Similarly, when a player moves down in rank, the sum of his directed rank differences increases by  $W_d$  and decreases by  $L_d$ . For any adjacent pair of players, A and B, where A is the higher ranked, if

$$W_d(A) - L_d(A) < W_d(B) - L_d(B), \tag{3}$$

a swap will produce an overall decrease in directed rank difference. By an appropriate sequence of swaps, a ranking can be produced that minimizes the algebraic sum of directed rank differences. This is perhaps best seen by working out an actual case.

A statistic related to  $W_d - L_d$  is *dominance percentage*  $P_d$ , defined as the percentage of the playing field dominated by the player in question, plus half the percentage which stands in no dominance relation with that player (either

---

<sup>9</sup>GJ

dominant or subordinate). It is convenient to think of dominance percentage as the percentage score that would result from a single game against other players in the competing field, assuming a draw against any player not actually encountered. If  $n$  is the total number of potential opponents in the field,

$$P_d = \frac{W_d + .5D_d}{n}, \quad (4)$$

where  $D_d$  is the number of opponents who do not stand in a dominance relation to the player in question.

$W_d - L_d$  is a linear transformation of  $P_d$ . From the definition of dominance percentage,

$$W_d = nP_d - .5D_d, \quad (5)$$

and by complement,

$$L_d = n - W_d - D_d. \quad (6)$$

Substituting (5) into (6) ,

$$L_d = n - nP_d - .5D_d, \quad (7)$$

and subtracting (7) from (5),

$$W_d - L_d = 2n(P_d) - n, \quad (8)$$

which is the linear transformation claimed. Sorting the field by  $P_d$  is therefore equivalent to sorting the field by  $W_d - L_d$  and has the same effect in minimizing directed rank differences.

Sorting by  $W_d - L_d$  or  $P_d$  does not minimize violations but is a useful algorithm for reducing them. In his master's thesis the author pitted this algorithm against a brute-force method in a computer simulation. The downloadable version produces a sequence of 100 all-play-all tournaments, each consisting of ten players (see "Downloads"). The players are all of equal playing strength inasmuch as their odds of winning in any pairing are equal. First, for each tournament the players are ranked by the dominance statistics developed here, using recursive tiebreakers where possible. Second, using the same outcomes, a ranking for each tournament is found by a brute-force method that minimizes violations. The rankings by dominance yielded a mean of 11.75 over the 100 tournaments, while the rankings by brute force yielded a mean of 9.3. The reader may judge whether the minimization of violations was worth the application of an onerous brute-force method. For larger tournaments the brute-force algorithm becomes impractical, in which case ranking by dominance statistics is the only recourse. As a demonstration of persistence, the author further showed ranking by dominance statistics to be more efficient in reducing total violations than some sophisticated probability algorithms of the day.<sup>10</sup>

The principle of minimizing directed rank differences explains the efficiency of an all-play-all tournament, which has long been regarded as a rational means

---

<sup>10</sup>J, 2.2

for ordering performances. The percentage scores resulting from such a tournament, including half points for draws, are equivalent to dominance percentages, and their descending order consequently minimizes directed rank differences. The principle may tentatively be extended to partial tournaments. An interesting illustration is the Mannheim 1914 tournament, an 18-player round-robin that was abandoned after 11 rounds with the outbreak of the First World War.<sup>11</sup> Alekhine then had the best score with 9.5 points. It has been argued that Vidmar with 8.5 points actually outperformed Alekhine because he had met stronger opposition, which is supported by several auxiliary scoring methods, including Neustadtl and Solkoff. Sorting the field by  $P_d$  gives the nod to Alekhine. The dominance percentage for Alekhine was 12.5/17, and for Vidmar 11.5/17.

A more sophisticated treatment anticipates the discussion of interval ratings. For the moment we simply note that a simultaneous calculation of interval ratings ( $K = 1$ ) for the Mannheim data gives a rating of .3769 for Alekhine's performance, and a rating of .3994 for Vidmar's. (The Excel program, Mannheim1914 under the heading "Ordinal Ratings 2," may be downloaded from *ftp.rathingtontheory.com*.) This anomalous result is due to the fact that the data is from a partial tournament. For a complete (all-play-all) tournament performance can be calculated from the average pre-event ratings of the participants as

$$R = ER_a + K(P - P_c) \frac{M - 1}{M}. \quad (9)$$

for  $M$  participants. The average opposition rating, in effect, is the same for all the participants. The only variables with respect to individual ratings are the player's percentage score and its complement, the opposition percentage score. Consequently, percentage scores completely determine the ordering of performances. (Formula (9), along with its proof, was attributed by Elo to Nida Uzman, "young Turkish chess player and a mathematics student in Berline University."<sup>12</sup>)

The dominance percentage statistic provides a formula for ordinal ratings:

$$P = \text{rank}(P_d), \quad (10)$$

which is analogous to the basic formulas of the more advanced rating systems to be introduced under subsequent headings. Ordinal ratings, thus defined, minimize directed rank differences. We have already seen their limitations in the treatment of the Mannheim data.

Formula (10) can be applied all at once to a competing field. As with interval and ratio ratings, there is also a sequential algorithm. The algorithm assumes a competing field that has already been assigned ordinal ratings. For each player a list of opposition ratings is maintained in ascending order, say the ratings of the last  $n$  opponents that the player has encountered. Using the player's point score against the current list opponents, the player's ordinal rating is calculated.

<sup>11</sup>HW, "auxiliary scoring methods"

<sup>12</sup>E3

Suppose a player has in the list of  $n$  opposition ranks the values

$$R_1, R_2, R_3, \dots, R_{n-1}, R_n$$

in ascending order and that the player has scored  $s$  points against the  $n$  opponents. The player's ordinal rating would fall between  $R_s$  and  $R_{s+1}$ . The rating is then used in the subsequent calculation of opposition ratings, and the algorithm proceeds in self-corrective fashion.

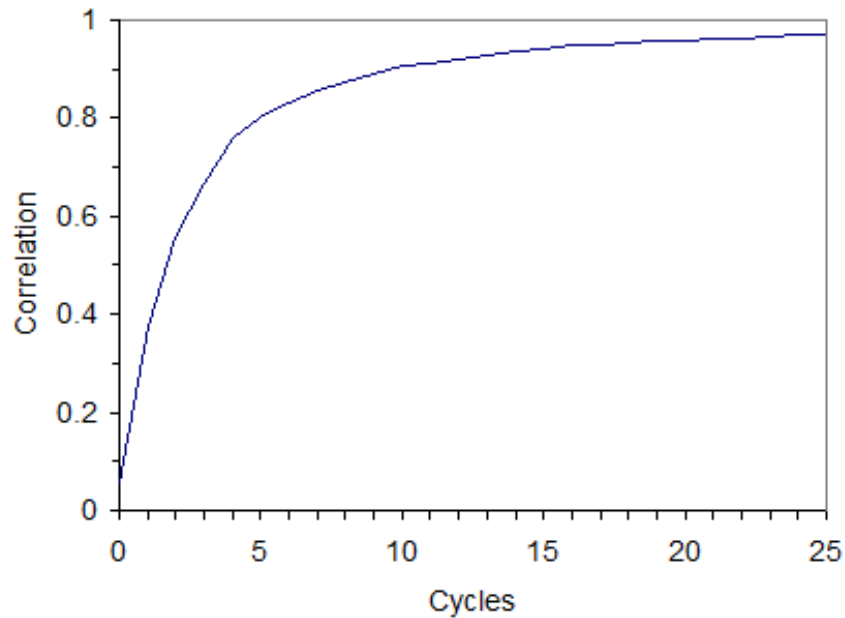


Figure 2: Correlation of Calculated Ranks with Predefined Ranks by Spearman's Rank Correlation Coefficient

This scheme is implemented in a program available by download from *ftp.RatingTheory.com* in the folder "Ordinal Ratings 3" (See "Downloads"). The program uses a ranking of 400 players in 25 cycles of 200 games each, but users may change the experimental parameters. The pre-assigned ranks represent the "true" playing strength of each player. Players are also assigned initial quantile values. Randomly distributed quantiles were found to work best. Pairings occur at random, and the optimistic assumption is made that the stronger player invariably wins. A small list of opposition quantiles, four for each player, proved adequate. At the end of each cycle the correlation between pre-assigned ranks and ranks based on quantiles was determined, as depicted in the graph above. The neat results are obviously due to the artificial conditions of the simulation, but the primary object was to demonstrate the feasibility of the system. It is

doubtful, of course, that such an algorithm could ever be more than a toy rating system.

## 4 Interval Ratings

A statistical theory of ratings begins with the discovery that a typical rating system, such as the Ingo, is tacitly treating of *differences* in percentage score. Consider a sequence of games between two players, *A* and *B*. If *A*'s percentage score in this sequence is  $P$ , then *B*'s score is  $1 - P$ , and the difference in percentage score  $P(A) - P(B)$  is

$$P - (1 - P) = 2P - 1 = 2(P - .50).$$

This last expression recalls the basic Ingo formula (1) and suggests its motivation. We can generalize this discovery by postulating a principle of interval systems: *differences in rating reflect differences in percentage score*. A rating formula that captures this principle is

$$R = ER_c + K(P - P_c). \tag{11}$$

The symbol *E* (expected) is used here to designate an arithmetic mean, so that  $ER_c$  represents the mean rating of opponents.  $K$  is an arbitrary constant (-50 in the Ingo System).  $P$  and  $P_c$  are the percentage scores of player and opposition. The difference may also be written as

$$\frac{W - L}{N} \tag{12}$$

for points won and lost out of  $N$  games. A term of convenience for the difference in percentage score is *relative performance*, which will later be broadened to include ratios.

The effect of (11) when applied to a competing field is to generate rating differences in proportion to relative performance, which is more easily seen by writing the formula as

$$R - ER_c = K(P - P_c). \tag{13}$$

The latter may be viewed as an equation of means over game instances,

$$E[R - R_c] = E[K(S - S_c)], \tag{14}$$

where the relative score,  $S - S_c$ , in chess evaluates to 1, -1, or 0. For individual games, rating difference may be thought of as predicting relative score as an approximation, and the question naturally arises as to how good this approximation is. The efficiency of linear rating systems relies on a basic statistical argument: Simply put, the mean of a distribution is the value that minimizes the sum of squared deviations of the scores. This can be demonstrated by the mathematically trained as an exercise in differential calculus, but the idea is basic enough to be taken for granted. We shall call the argument Gauss' principle since he appears to have been the first to have used it.

If the rating of Formula (11) is represented as the mean

$$R = E[R_c + K(S - S_c)], \tag{15}$$

it follows directly from Gauss's principle that

$$\sum (R - [R_c + K(S - S_c)])^2$$

is an absolute minimum over real values of  $R$  when (11) holds true. We have only to regroup terms as

$$\sum [(R - R_c) - K(S - S_c)]^2$$

to show that difference in rating predicts relative score in this least squares sense. The approximation is optimal for ratings calculated by the general linear formula, regardless of the consistency of data on which the ratings are based.

The foregoing argument would seem to be conclusive in favor of linear ratings, but since Elo's characterization continues to hold sway, some elucidation is called for. For of all, what is meant by the statement that relative score is predicted by rating difference in individual games? Imagine a set of linear ratings that have been calculated for a set of results in a competing field. We shall suppose for the sake of simplicity that all the results are either wins or losses. We shall further suppose that the ratings are consistent, that is, that they are all calculated from the same set of ratings. The latter condition, as we shall see, is not easily achievable in practice, but for now we simply state it as given. You, as arbiter in this issue, are given only one piece of data from this collection, namely a result that occurred between players  $X$  and  $Y$ , perhaps one of many, with  $X$  rated  $R$  and  $Y$  rated  $R_c$ . You are asked to guess whether the result was a win or a loss for player  $X$ . By manipulating (11) you can easily determine that the percentage score between Player  $X$  and players rated the same as  $Y$  was some value  $P$  from the view point of  $X$ . If  $P$  is greater than .5, your best guess would be that  $X$  won. If  $P$  is less than .5, your best guess is that  $X$  lost.

The prediction you make is a statistical one. It assumes nothing about the playing strength of the players in the competing field. For all you know, they could be a bunch of machines programmed to make random moves. Now suppose that you are asked to predict a further result between  $X$  and  $Y$  beyond the data given. If the players really are machines like the ones described, you might as well toss a coin. It is conceivable, on the other hand, that by continuing the competition indefinitely the percentage scores would tend toward fixed values as limits. The original scores would then represent estimates of their limiting values or probabilities, and the competing field would qualify as a *collective* in the sense used by Richard von Mises in his theory of probability.<sup>13</sup> Your further prediction would then be plausible.

Linear rating systems, to conclude, involve predictions in several senses of the term, and the probability issues that are raised are by no means trivial. The persistence of a popular rating system, however flawed, is therefore not to be underestimated.

---

<sup>13</sup>V

## 5 Ratio Ratings

Relative performance may be represented by a *ratio* of percentage scores as well as by a difference. The corresponding principle for ratio systems is that *rating ratios reflect percentage score ratios*, and the formula corresponding to (11) is

$$R = ER_c \frac{P}{P_c}, \quad P_c > 0, \quad (16)$$

$$= ER_c \frac{P}{1-P}, \quad P < 1, \quad (17)$$

$$= ER_c \frac{W}{L}, \quad L > 0, \quad (18)$$

with the same variables as the interval formula. Here relative performance cannot be defined for individual games because of the possibility of division by zero, and it must therefore be treated as a constant. Moving  $R$  to the right and relative performance to the left and reversing the equation,

$$\frac{ER_c}{R} = \frac{L}{W}, \quad (19)$$

which may also be written

$$E \frac{R_c}{R} = \frac{L}{W}. \quad (20)$$

We now have an equation of means to which Gauss' principle can be applied.

$$\sum \left( \frac{R_c}{R} - \frac{L}{W} \right)^2$$

is a minimum over real values of  $R$  for the calculated value. Rating ratios in individual games may thus be said to predict overall relative performance. The reciprocals in this result can be eliminated by using a harmonic mean in (16), but this would be creating practical problems to avoid a theoretical quirk.

Elo regarded ratio systems as an important advance over interval systems, perhaps because a ratio scale for physical measurement has distinct advantages. The two rating systems are distinguished primarily by the fact that aggregate scores can be related either as ratios or differences. When logarithms are employed to deal with the large values of a ratio system, the distinction becomes somewhat blurred, and it is not altogether clear that a ratio system is to be preferred.

Elo's focus in his ratio system was on established ratings. Performance ratings were left to be estimated by linear processes, but a performance rating can be derived from his formula (46),

$$P(D) = \frac{1}{1 + 10^{-D/2C}}. \quad (21)$$

Solving for  $D$  gives

$$D(P) = 2C \cdot \log_{10} \frac{P}{P_c}, \quad (22)$$

where  $C$  is the class interval of 200 rating points. A performance formula would therefore be

$$R = ER_c + 400 \cdot \log_{10} \frac{P}{P_c}, \quad P > 0, P_c > 0. \quad (23)$$

The conditions on this formula are no doubt reason enough to avoid it. The formula follows directly from (16) by the application of logarithms, although its use of arithmetic averaging implies a geometric average in the original formula. There is no special notation for ratings as logarithms. The base of logarithms  $b$  may be calculated from the coefficient of relative performance,

$$\log_b R = 400 \cdot \log_{10} R, \quad (24)$$

and by the rules of logarithms

$$\log_b R = \frac{\log_{10} R}{\log_{10} b}. \quad (25)$$

Setting equals to equal,

$$\frac{\log_{10} R}{\log_{10} b} = 400 \cdot \log_{10} R. \quad (26)$$

Consequently,

$$\log_{10} b = \frac{1}{400} \quad (27)$$

and

$$b = 10^{1/400} \quad (28)$$

or about 1.00577. Aside from inaccuracies arising from averaging, Elo's ratio system appears to be a logarithmic version of the system described here, and this confluence of ideas is telling. It suggests that reliance on a probability distribution based on the logistic function is unnecessary. A performance formula derived from (21), as noted, is not actually used in the system. Established ratings do the main work, and their formulas make assumptions about percentage expectancy that are characteristic of the Elo System. We shall postpone definite conclusions till the discussion of established ratings. It is perhaps worth noting that taking the logarithm of (16) does not produce a linear formula equivalent to (11) since relative performance in the result would be the nonlinear

$$\log_b P - \log_b P_c.$$

Another ratio system, one with unique properties, is the Berkin System, which will be discussed under that heading.

## 6 Sequential vs. Simultaneous Ratings

Rating systems in chess typically maintain a pool of ratings and apply their formulas to contests, either individual games or tournaments, as they occur among the rated players. This straightforward approach, producing *sequential ratings*, is not without its problems.

- Although the emphasis is usually on up-to-date ratings, the underlying data are an assortment of old and new.
- Data samples vary in size, from the few games of the occasional player to the many games of the enthusiast, with resulting variation in sampling error.
- Ratings may be manipulated by players using various methods, such as “sandbagging.”
- While some ratings are more or less stable, others are rising rapidly, and their interaction causes deflation in the pool at large.
- Finally, the rating pool itself is changing as players come and go.

The minimizing effect of the rating process will eventually make itself felt in a sequential system, but the effect on individual ratings in the meantime could be calamitous. The Elo System attempts to keep ratings up-to-date by limiting sample size in its established rating, which becomes a kind of moving average, described more fully under “Established Ratings.” Unlike ordinary moving averages, where sample size is restricted to the last  $N$  games, established ratings are based on attenuated sample weight. The effect of a rated game, having an original sample weight of  $1/N$ , becomes attenuated as more and more data are processed. In theory at least, the effect is never completely lost. The established rating becomes what might be called a *weighted moving average*. Although recent data are more heavily weighted, rating changes emerging from the averaging process do not keep pace with changes in playing strength. Timely adjustments, in short, do not guarantee currency of the data on which they are based.

The disadvantages of sequential ratings may be overcome by applying simultaneous calculations to a defined data set, using either matrix algebra or an iterative process. In 1969 such an application produced the first International Rating List, using a computer to make iterative calculations on the complete interplay of 210 contestants over the previous three years.<sup>14</sup> The resulting list was a self-consistent set of ratings with clearly defined boundaries for its data, which could then be used for sequential calculations. The downside of this approach is that for large systems it requires enormous computing resources, but with the computer power now available it is the rating method of choice.

There is a simple modification of sequential ratings that seems intuitively to be a first step towards simultaneous calculations. Notice that rating adjustments between pairs of contestants are based on pre-event ratings, so that the adjustment of a rating is mirrored in the opponent’s rating. Since two adjustments are redundant it seems logical to halve each one each one for linear ratings. This suggests an experiment in which the efficiency of linear adjustments for a defined set of outcomes is compared to the efficiency of the same adjustments by half. Efficiency in this context refers to correlations between rating differences and relative performance in the form of score differences,  $W - L$ . The rough

---

<sup>14</sup>E2, Part 3

and ready simulation summarized in the graph below performs this experiment on a ranking of 100 players for 800 random pairings in stages of 100 (“Downloads”). Results are rated sequentially with the higher ranked player invariably winning. After each stage an error statistic is calculated for all pairings in the field, namely, the root mean square of the deviation of relative performance,  $W - L$ , from expected relative performance based on the generated ratings. Initially the error statistic is lower for the half adjustments, but this advantage gradually disappears with the increase of pairings. It seems, then, that this approach to simultaneous calculations has at best a temporary benefit.

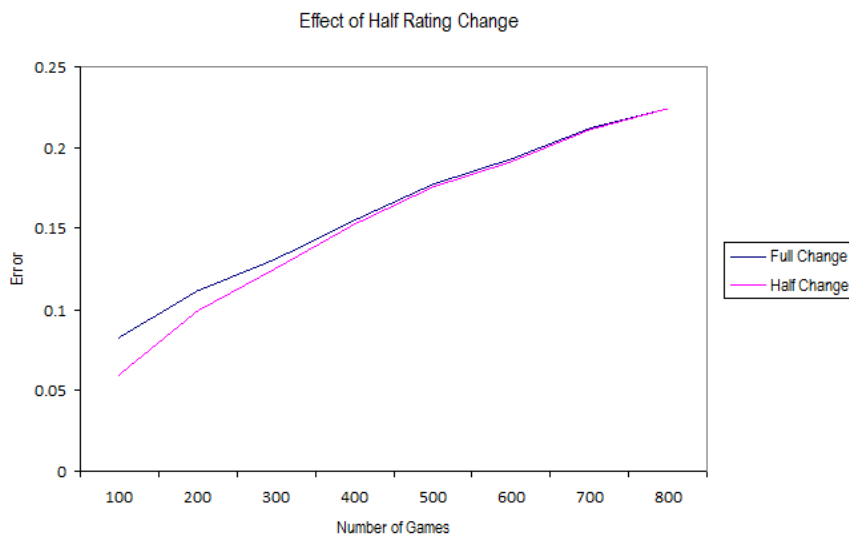


Figure 3: Effect of Half Rating Change

## 7 Consistency

Ratings are typically calculated event by event in sequential fashion, but there are distinct advantages to calculating them simultaneously over a defined period of time for a defined competing field. The simultaneous approach, while considerably more onerous, generates ratings that are mathematically consistent. The implicit assumption of sequential ratings is that the active rating pool will eventually reach a similar state of consistency, but this is hardly more than wishful thinking. Elo’s terminology with regard to this distinction can be confusing. He essentially divided sequential ratings into two types: *continuous* ratings, which are calculated event by event, and *periodic* ratings, which are calculated for

calendar periods.<sup>15</sup> For the first International Rating List (see “Sequential vs. Simultaneous Ratings”) he employed simultaneous ratings in the form of iterative calculations on a computer. Calculations for this list were “continued until successive values of the differences showed little or no significant change,” eight iterations in all.<sup>16</sup> This produced a more or less self-consistent set of ratings.

Consistency for a small data set can be studied with matrix manipulation. Let us test the systems under discussion using the hypothetical tournament below.

Table 1: Single Round Robin for Four Players

Players	A	B	C	D	wins	pct.
<b>A</b>	x	1	1	0	2	.667
<b>B</b>	0	x	1	1	2	.667
<b>C</b>	0	0	x	1/2	.5	.167
<b>D</b>	1	0	1/2	x	1.5	.500

As seen from its matrix representation, the system of linear formulas for this tournament has an infinite number of solutions, with the rating of player D as a free variable (Tables 2 and 3). A unique solution is reached by assigning D a rating and calculating the other ratings accordingly. For example, with D rated .5, the solution is (.75, .75, 0, .5). Systems of ratio formulas, including the

Table 2: Matrix Representation of Table 1: Formula (11) ( $K = 1$ )

	A	B	C	D	
<b>A</b>	1	-1/3	-1/3	-1/3	1/3
<b>B</b>	-1/3	1	-1/3	-1/3	1/3
<b>C</b>	-1/3	-1/3	1	-1/3	-2/3
<b>D</b>	-1/3	-1/3	-1/3	1	0

Table 3: Row Canonical Form for Table 2

	A	B	C	D	
<b>A</b>	1	0	0	-1	1/4
<b>B</b>	0	1	0	-1	1/4
<b>C</b>	0	0	1	-1	-1/2
<b>D</b>	0	0	0	0	0

Berkin formula (79), are homogeneous, allowing only the zero solution (0, 0, 0, 0). For basic ratio ratings (16) there is a strategy for finding nonzero solutions where there is an undefeated player, whose rating is ordinarily undefined because

<sup>15</sup>E1, 1.5-6

<sup>16</sup>E2, Part 3

of division by zero. Consider, for example, the addition of Player E to the above round robin with a single win against Player A (Table 4). A matrix for the corresponding ratio formulas, with player E assigned an arbitrary rating of 1, is given in Table 5. A solution is provided by mathematical software as (.6571, .9143, .1429, .5714). There does not seem to be a corresponding strategy for Berkin ratings, but the overdetermined matrix of Table 6 gives the solution (.3684, .3158, .0526, .2632) with the added convenience of assigning an arbitrary mean (here .25). Finally, there is the matrix representation of the system of Elo formulas, calculated by Formula (23) (Table 7). Mathematical software indicates no solution, and manipulation does not proceed beyond the echelon form of Table 8.

Table 4: Round Robin of Table 1 with Additional Result

Players	A	B	C	D	E	wins	pct.
A	x	1	1	0	0	2	.500
B	0	x	1	1	x	2	.667
C	0	0	x	1/2	x	.5	.167
D	1	0	1/2	x	x	1.5	.500
E	1	x	x	x	x	1	1.000

Table 5: Matrix Representation of Table 4: Formula (16)

	A	B	C	D	
A	1	-1/4	-1/4	-1/4	1/4
B	-2/3	1	-2/3	-2/3	0
C	-1/15	-1/15	1	-1/15	0
D	-1/3	-1/3	-1/3	1	0

Table 6: Matrix Representation of Table 1 with Total Ratings: Formula (79)

	A	B	C	D	
A	1	-1	-1	0	0
B	0	1	-1	-1	0
C	0	0	1	-1/5	0
D	-2/3	0	-1/3	1	0
$\Sigma$	1	1	1	1	1

These results are confirmed by iterative calculations on the all-play-all results of Table 1 starting with a single arbitrary rating. As seen in Table 9, ratings converge rapidly to their final values by the linear formula. The Berkin formula also shows convergence, though at a slower rate. Neither the Elo formula nor the general ratio formula yields convergent ratings.

Table 7: Matrix Representation of Table 1: Formula (23)

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	
<b>A</b>	3	-1	-1	-1	361.236
<b>B</b>	-1	3	-1	-1	361.236
<b>C</b>	-1	-1	3	-1	-838.764
<b>D</b>	-1	-1	-1	3	0

Table 8: Echelon Form for Table 7

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	
<b>A</b>	3	-1	-1	-1	361.236
<b>B</b>	0	8	-4	-4	1444.944
<b>C</b>	0	0	48	-48	-11460.672
<b>D</b>	0	0	0	0	-133968.384

Large-scale applications of simultaneous ratings require iterative calculations, such as were used by Elo in compiling the initial International Rating List. A similar application was tried by this author using Microsoft Excel spreadsheets to see whether available commercial software could handle the task (See “Downloads”). The spreadsheets exploit Excel’s handling of “circular references,” normally considered errors, to produce self-consistent ratings. The mode of calculation has been set to manual, and the “calculate on save” option has been turned off. Initially only the results of the first cycle of calculations are shown. The buttons under Calculation on the Formulas tab (or F9) produce subsequent cycles.

The spreadsheet for simultaneous linear ratings (CrossTable2007) uses 265 USCF-rated games played by the 42 members of the Cranston-Warwick Chess Club (RI) in the calendar year 2007. Again, the function key F9 causes the ratings to converge in stepwise fashion. To the far right of the spreadsheet, ratings are rendered in more familiar formats. Format 1 is based on a fractional scale from 0 to 1. The original ratings  $R$  are converted using minimum and maximum values, as

$$R_f = \frac{R - R_{min}}{R_{max} - R_{min}}. \tag{29}$$

The resulting fractional rating is converted in turn to a kind of Elo rating based on a linear scale (Format 2):

$$R_E = Elo_{min} + R_f(Elo_{max} - Elo_{min}). \tag{30}$$

Since the high Elo rating in the club was close to 2000, and the low Elo rating was close to 1000, it was decided to use the interval 1000 to 2000 for the unofficial 2007 Elo club ratings.

Another spreadsheet available by download (Champs2008) is for simultaneous Berkin ratings, using data from the 2008 club championship of the Cranston-

Table 9: Convergence of Recursively Calculated Ratings  
In a Single Round Robin (Table 1)

<b>Linear Formula</b> (11), K = 1, values initialized at .5				
	A	B	C	D
Iteration 5	0.751029	0.751029	-0.00205761	0.5
6	0.749657	0.749657	0.000685871	0.5
7	0.750114	0.750114	-0.000228624	0.5
8	0.749962	0.749962	7.62079e-005	0.5
9	0.750013	0.750013	-2.54026e-005	0.5
10	0.749996	0.749996	8.46754e-006	0.5
11	0.750001	0.750001	-2.82251e-006	0.5
12	0.75	0.75	9.40838e-007	0.5
13	0.75	0.75	-3.13613e-007	0.5

<b>Berkin Formula</b> (79), values initialized at .5				
	A	B	C	D
Iteration 50	0.913513	0.783039	0.130499	0.651468
60	0.913294	0.782406	0.130387	0.652222
70	0.912974	0.782585	0.130433	0.652239
80	0.913025	0.782631	0.13044	0.652163
90	0.913052	0.782609	0.130435	0.652168
100	0.913045	0.782606	0.130434	0.652175
110	0.913043	0.782609	0.130435	0.652174
120	0.913043	0.782609	0.130435	0.652174

<b>Elo Formula</b> (23), values initialized at 2200				
	A	B	C	D
Iteration 100	1328.48	1328.48	1028.48	1238.17
110	1231.57	1231.57	931.567	1141.26
120	1134.66	1134.66	834.657	1044.35
130	1037.75	1037.75	737.747	947.438
140	940.837	940.837	640.837	850.528
150	843.927	843.927	543.927	753.618

<b>Ratio Formula</b> (16), values initialized at 1				
	A	B	C	D
Iteration 10	7.80455	7.80455	1.15829	4.76605
20	38.3958	38.3958	5.69836	23.4469
30	188.893	188.893	28.0339	115.350
40	929.287	929.287	137.917	567.483
50	4571.75	4571.75	678.499	2791.81
60	22491.4	22491.4	3337.97	13734.7

Warwick Chess Club. Convergence by the function key F9 in this case is quite rapid. On the far right of the spreadsheet, the ratings are rendered as natural logarithms and as pseudo-Elo ratings. The Berkin ratings were initialized so as to produce plausible Elo ratings as a function of 400 times the common logarithm.

## 8 Established Ratings

Established ratings are an alternate method of calculation replacing performance ratings when the latter become unwieldy. To understand them it is necessary to begin with the performance formula, which Elo gives as his first formula,

$$R_p = R_c + D(P). \quad (31)$$

“ $D(P)$  is to be read as the difference based on the percentage score  $P$ , which is obtained from the curve or table.”<sup>17</sup> We have seen in the discussion of probability that the normal curve is a questionable basis for  $D(P)$  in the interval system, and in the ratio system it is not clear that  $D(P)$  is a probability function. We proceed, in any case, to the established rating.

Established ratings are calculated by a change formula arising from the mathematical procedure known as *cumulative averaging*. We have seen from the discussion of interval ratings that performance ratings may be treated as averages. Given an old rating  $R_o$  based on  $N_o$  games and a performance rating  $R$  based on  $N$  games, the ratings are combined as

$$R_n = \frac{R_o N_o + RN}{N_o + N}. \quad (32)$$

The old rating may be represented by the identity

$$R_o = \frac{R_o N_o + R_o N}{N_o + N}. \quad (33)$$

Subtracting  $R_o$  from  $R_n$  gives

$$\Delta R = \frac{(R - R_o)N}{N_o + N}, \quad (34)$$

and repeating this process gives a recursive change formula for cumulative averaging. Rather than allowing the old sample to increase indefinitely, we can maintain it at a constant value. This we call the *sampling constant* to distinguish it from ordinary samples. The resulting error in cumulative averaging is negligible if the constant is sufficiently large. If we substitute  $N_o - N$  for  $N_o$  in (32) we get

$$R_n = \frac{R_o(N_o - N) + RN}{N_o}, \quad (35)$$

---

<sup>17</sup>E1, 1.51

which is Elo's "blending process".<sup>18</sup> Writing the original rating  $R_o$  as the identity

$$R_o = \frac{R_o(N_o - N) + R_o N}{N_o}, \quad (36)$$

a change formula follows by subtraction as

$$\Delta R = (R - R_o) \frac{N}{N_o}, \quad N < N_o. \quad (37)$$

A more accurate procedure is to hold  $N_o$  constant in (34), which puts no restraint on  $N$ . The rating change thus far has been a straightforward application of cumulative averaging, but at this point complications arise in Elo's procedure. The rating difference  $R - R_o$  becomes a change in rating difference by subtracting the old performance rating,

$$R_o = R_c + D(P_o), \quad (38)$$

from the new performance rating,

$$R_p = R_c + D(P), \quad (39)$$

and "making the simplifying assumption that  $R_c$  is the same for both samples,"

$$\Delta R = [D(P) - D(P_o)] \frac{N}{N_o}. \quad (40)$$

The difference in  $D$  represents the change in rating, but it is the difference in percentage score that is "read from the percentage expectancy curve, with slope  $S$ " to get the rating change, which presumably means that the derivative of the curve is applied to the percentage difference. The entire procedure, beginning with (37), is somewhat roundabout. It is clear from (39) that  $P$  is the percentage score encountered in the new performance, and from (38) that  $P_o$  can be calculated from (31) using the new opposition rating. The difference in percentage score can then be translated into a rating change using the derivative of the performance rating with respect to  $P$ , which in this case amounts to the inverse slope of the Percentage Expectancy Curve. The established formula becomes

$$\Delta R = R'(P - P_o) \frac{N}{N_o}, \quad (41)$$

where  $R'$  is the derivative of the basic rating formula with respect to  $P$ . The result is the same, but the new procedure offers the possibility of dispensing with the probability distribution altogether. It furthermore suggests a new interpretation of the established rating. Since the change in percentage score arises from cumulative averaging, it may be regarded as itself the result of cumulative averaging. On the assumption that percentage scores tend to long-term limiting values, the old percentage score represents an expected value which

---

<sup>18</sup>E1, 8.63, described in 8.25

is corrected by the additional score. For nonlinear ratings in general, formula (41) becomes a differential yielding an approximate value for rating change where the difference in percentage score is small. The differential simplifies by multiplication to

$$\Delta R = \frac{R'(W - W_e)}{N_o}. \quad (42)$$

A constant  $K$  customarily combines the derivative and the sampling constant, giving

$$\Delta R = K(W - W_e), \quad (43)$$

which is Elo's established rating formula. When  $D(P)$  in (31) is based on the Percentage Expectancy Curve,  $P(D)$ , the derivative  $R'$  is taken from the inverse function. The derivative of the curve is approximated as the average slope of "the most used portion," roughly 1 percentage point over 8 points of rating difference.<sup>19</sup>  $R'$  is the reciprocal of this approximation, 800 rating points. This simplified derivative applies to either the normal curve or the logistic function, and the complicated Verhulst distribution, Elo's formula (45), goes by the board.

The sampling constant in the Elo System,  $N_o$ , is taken to be the maximum sample size for the performance rating, which is then balanced against the size of a further performance sample,  $N$ , in the established rating. To establish an appropriate value for the sampling constant, Elo compared  $N_o$  to  $N$  with respect to sampling error.<sup>20</sup> This analysis overlooks the recursive nature of the process, which is somewhat obscured by the notation. Assuming for the sake of simplicity events of equal size  $N$ , a performance that starts with a sample weight of

$$\frac{N}{N_o}$$

has a sample weight after  $q$  calculations of

$$\frac{N}{N_o} \left(1 - \frac{N}{N_o}\right)^{q-1}.$$

The sample weight of an event never quite disappears as the number of events on which the established rating is based becomes indefinitely large. The formula is plotted below for some typical values of sampling weight, with  $N = 5$ .

## 9 Attenuation

Attenuation is a byproduct of Elo's handling of cumulative averaging:

There are several combination processes. One might simply average the results from  $N$  and  $N_o$ , obtaining a new average rating for the entire sample  $N_o + N$ , but this preserves fully the rating contribution of the earlier samples and produces, if real changes in ability have occurred, a false statement of the current rating.

---

<sup>19</sup>E1, 8.25

<sup>20</sup>E1, 8.28

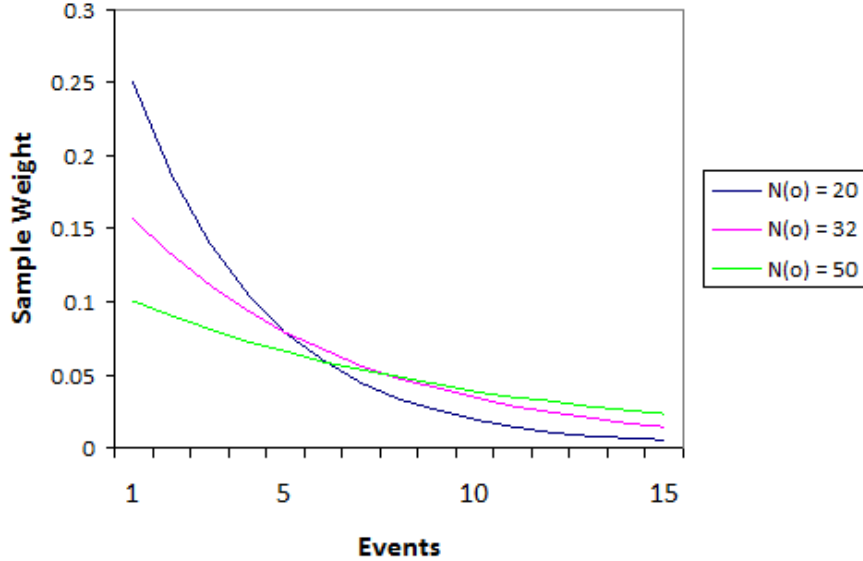


Figure 4: Attenuation of Sample Weight

The proper method to combine  $R_p$  with  $R_o$  should attenuate the earlier performances in favor of the later ones . . . <sup>21</sup>

Formula (34), which uses the entire sample  $N_o + N$ , does not fully preserve the contribution of earlier samples because  $N_o$  is maintained as a constant from one event to the next, but it is more accurate as an average than Elo's blending process. The sample weight of the original rating in the blending process,

$$\frac{(N_o - N)}{N_o},$$

clearly depends on the size of the new sample, which must never exceed the size of the original. The process does produce attenuated samples, as depicted in the graph of the previous section.

In the Elo System attenuation does not occur until the transition to established ratings, but there is no logical reason why it should not occur from the outset of the rating process. Attenuation can be achieved directly by an attenuating factor  $\alpha$ , such that  $0 < \alpha < 1$ , which is applied to a sequence of change formulas for cumulative averaging. It will be helpful first to have an accurate description of cumulative averaging. The process requires that separate records be maintained for the average itself, in this case the cumulative rating, and for

---

<sup>21</sup>E1, 8.25

the sample on which it is based, which is also cumulative. This gives the two formulas

$$R_{i+1} = \frac{R_i N_i + RN}{N_{i+1}} \quad (44)$$

and

$$N_{i+1} = N_i + N. \quad (45)$$

We use a simple notation in which subscripted variables refer to cumulative values, while unsubscripted variables refer to incremental values, i.e., a new performance rating and its sample size. The next step in the cumulative process would be

$$R_{i+2} = \frac{R_{i+1} N_{i+1} + RN}{N_{i+2}} \quad (46)$$

and

$$N_{i+2} = N_{i+1} + N. \quad (47)$$

The unsubscripted values,  $R$  and  $N$ , have most likely changed since they refer to a new performance. The whole process begins with a single performance rating, as

$$R_0 = R, \quad (48)$$

and

$$N_0 = N. \quad (49)$$

Note that cumulative results are no different from those that would be obtained by ordinary averaging, given the same data. The difference is that in cumulative averaging we do not have to write out the the results of all the previous steps. Now we apply the factor  $\alpha$  to obtain attenuated ratings.

$$R_{i+1} = \frac{\alpha R_i N_i + RN}{N_{i+1}} \quad (50)$$

and

$$N_{i+1} = \alpha N_i + N \quad (51)$$

for  $i = 0, 1, 2, 3, \dots$

Though difficult to describe succinctly, the method is clearly superior to those that rely on a sampling constant. It is based on true cumulative averaging and can be applied from the outset of the averaging process. Its primary disadvantage, apart from the fact that it is still a sequential process, is that a record of sample size must be maintained in addition to the rating.

As an interesting sidelight, if we write the expansions for a series of attenuated sample sizes,

$$N_o = N, \quad (52)$$

$$N_1 = \alpha N_o + N,$$

$$N_2 = \alpha^2 N_o + \alpha N_1 + N,$$

$$N_3 = \alpha^3 N_o + \alpha^2 N_1 + \alpha N_2 + N,$$

and so forth, we get a geometric series. Assuming  $N$  to be constant,

$$\lim N_i = \frac{N}{1 - \alpha}$$

as  $i$  goes to infinity. Now if

$$\alpha = \frac{N_o - N}{N_o},$$

the weight of the old sample in Elo's blending process, then the limit is the sampling constant  $N_o$ . Furthermore, if

$$\alpha = \frac{N_o}{N_o + N},$$

the weight of the old sample in formula (32), then the limit is  $N_o + N$ .

## 10 Percentage Expectancy

Attempts to apply probability theory to the rating process must deal with two equally plausible definitions of probability. The first definition invokes a sample space divided into  $n$  subsets of equally likely outcomes. If an event is associated with  $r$  of these outcomes, then its probability is  $r/n$ . This definition is at the heart of probability distributions, such as the normal or logistic. The Percentage Expectancy Curve, which is patterned after these distributions, is not really a probability distribution in the usual sense. Instead, it arbitrarily assigns probability values to rating differences in imitation of these important distributions. In an actual distribution of rating differences, the differences above or below a certain percentile tell us nothing about percentage expectancies.

The second definition of probability invokes the long-term limit of the relative frequency of an event. If an event occurs  $r$  out of  $n$  times as  $n$  goes to infinity, then its probability is  $r/n$ . The percentage scores encountered in rating systems represent estimates of this probability relative to rating differences or ratios. Rating systems are conservative in assuming, even in the face of evident changes in playing strength, that percentage scores tend toward a long-term limit. For a given pair of ratings, percentage expectancy is the hypothetical result that produces no change in the ratings. It is calculated in any rating system as the inverse of its basic formula, solving for percentage score.

The development of Elo's ratio system is very different from that of his interval system, but the emphasis in each case is on a probability distribution. We have seen that there are problems with deriving percentage expectancies from the normal curve. It is possible that unease with the derivation led Elo to his ratio system, both to set out in a new direction and to confirm his earlier work. In any case, he went on to develop a ratio version of his rating system which was eventually implemented by the USCF and FIDE. The development of the system begins with a disarmingly simple premise. Given the odds of Player

$x$  defeating Player  $y$  and the odds of Player  $y$  defeating Player  $z$ , then the odds of Player  $x$  defeating Player  $z$  are

$$\frac{P_{xy}}{P_{yx}} \cdot \frac{P_{yz}}{P_{zy}} = \frac{P_{xz}}{P_{zx}} \quad (53)$$

where  $P_{xy}$  is the probability of  $x$  scoring over  $y$ , etc.<sup>22</sup> This neat little result should immediately raise red flags. Although the result is given in the form of odds, it implies that if we know the probabilities of  $x$  defeating  $y$  and of  $y$  defeating  $z$ , then the probability of  $x$  defeating  $z$  follows as a logical consequence. We know from experience that because  $x$  defeats  $y$  and  $y$  defeats  $z$ , it is by no means certain that  $x$  will defeat  $z$ . Does expressing this in terms of odds make the observation any more certain? Formula (53) posits a transitive relationship among the probabilities. An amusing counterexample was provided by Martin Gardner (1914-2010) in one of his mathematical sketches.<sup>23</sup> He reported on a set of four dice cleverly constructed to demonstrate nontransitive probabilities. In a game that makes use of these dice, the first player selects a die with the idea of maximizing his chances in a roll-off against the second player. The second player is always able to select one of the remaining dice such that the odds of winning the roll-off are 2:1 in his favor. It would seem that there must be one die among the four, call it  $x$ , such that for any of the three remaining dice, call it  $y$ , the odds of the first player winning with  $x$  against  $y$  are at least even, but this is not the case. This is because the probabilities involved are not transitive. Before accepting (53) at face value, we are inclined to doubt the propriety of multiplying odds in this fashion.

Formula (53) does not figure directly in the development of the logistic function, the basis of Elo's ratio system. It does affect our interpretation of the logistic function as a probability distribution based on what Elo called a *logarithmic interval scale*. The development of the logistic function begins with his formula (39), which happens to be equivalent to our formula (16). It is not surprising, then, that we can derive the logistic function from the latter. Solving (16) for  $P$  gives a formula for percentage expectancy as

$$P_e = \frac{R}{R + ER_c}. \quad (54)$$

Dropping the mean indication, in a logarithmic system

$$P_e = \frac{b^R}{b^R + b^{R_c}} \quad (55)$$

for its base  $b$ . In Elo's logistic system, as we saw in "Ratio Systems,"

$$b = 10^{1/400}. \quad (56)$$

Consequently,

$$P_e = \frac{10^{R/400}}{10^{R/400} + 10^{R_c/400}}. \quad (57)$$

---

<sup>22</sup>E1, 8.33

<sup>23</sup>G, "Nontransitive Dice and Other Paradoxes"

Dividing top and bottom by  $10^{R/400}$ ,

$$P_e = \frac{1}{1 + 10^{(R_c - R)/400}}. \quad (58)$$

Substituting the variables  $C = 200$  and  $D = R - R_c$ ,

$$P_e = \frac{1}{1 + 10^{-D/2C}}, \quad (59)$$

which is the logistic formula for Elo's Percentage Expectancy Curve.<sup>24</sup>

A ratio system has the advantage of an unlimited rating scale, although the extremes of the scale are of limited interest. It is sometimes objected that the zero point on an interval scale is unrealistic because an upset in any pairing is possible. In theory, if an upset is *impossible*, then the probability of the weaker player winning is zero, but the converse is not true. By the frequency definition of probability, a probability of zero means a relative frequency that tends to zero as a limit, which does not exclude the possibility of an upset.

Elo's speculation that prolonged use of an interval system "draws the players in the pool together, eventually into a 4C range, filling out [a rectangular pattern]" is based on his view of rating formulas as probability distributions. There is a tendency, as Elo himself noted, for averaging to counteract the effect by the Central Limit Theorem.<sup>25</sup>

## 11 A Revised Elo System

We now come to the crux of our argument, which focuses on the interpretation of the established rating, represented here in abstract form as the differential

$$\Delta R = d\Delta P. \quad (60)$$

As Elo saw it, the established rating was essentially a small shift along the Percentage Expectancy Curve. The derivative  $d$  is based on the slope of the curve, and the change in rating is calculated from the curve as the rating difference associated with the change in percentage score. The interpretation offered here is quite different. The change in percentage score arises from cumulative averaging as a correction of the estimate of long-term percentage score. The correction is translated by the derivative  $d$  of the performance rating with respect to  $P$  into the rating change. The different interpretation, as it happens, does not affect practical results, but it does tend to rule out the need for the Percentage Expectancy Curve.

We shall illustrate these different points of view by deriving a new version of the established formula, and by implication a new rating system. The new

---

<sup>24</sup>E, 8.43

<sup>25</sup>E1, 8.57

system is logarithmic, like the Elo System, and uses the same constants. The formula for percentage expectancy is also the same as the one derived by Elo,

$$P_e = \frac{1}{1 + 10^{-D/400}}, \quad (61)$$

but not from a probability distribution. To avoid problems with averaging, we shall rate results one game at a time. The cumulative average for percentage score is consequently

$$\Delta P = \frac{S - P_e}{N_o}. \quad (62)$$

To translate this into a rating change we take the derivative of the performance formula (23) derived in our discussion of ratio ratings,

$$R' = \frac{400}{P(1 - P) \cdot \ln 10}. \quad (63)$$

The new established rating formula is therefore

$$\Delta R = \frac{400(S - P_e)}{P_e(1 - P_e) \cdot N_o \cdot \ln 10} \quad (64)$$

where  $P_e$  is defined in (61). The rating change resulting from a win in a pairing of various rating differences, positive and negative, is reported in the table below and compared with the rating change for the Elo System, interval and ratio. The rating changes resulting from a loss are obtained by changing  $S$  as well as the sign of  $D$ , giving negative versions of the reported values. The program used to generate the table is available by download (see “Downloads”). Since there are in theory no practical differences between the revised Elo and the logistic Elo, the different results are likely due to a more precise definition of the derivative in (63). Note that by taking the inverse of the derivative,

$$dP/dD = \frac{P(1 - P) \cdot \ln 10}{400}, \quad (65)$$

and substituting for  $P$  by the logistic formula (61), we get

$$dP/dD = \frac{(10^{-D/400}) \cdot \ln 10}{400(1 + 10^{-D/400})^2}, \quad (66)$$

which is Elo’s Verhulst formula (45).<sup>26</sup>

## 12 Cumulative Averaging and the Differential

The Elo System and its revised version postulate different uses of the derivative in arriving at the differentials known as established ratings. In the former the

---

<sup>26</sup>E1, 8.43

Table 10: Rating Change Produced by a Win for Various Rating Differences

	Revised Elo		Normal Elo		Logistic Elo	
	D>0	D<0	D>0	D<0	D>0	D<0
0	6.95	6.95	8	8	8	8
20	6.57	7.37	7.55	8.45	7.54	8.46
40	6.23	7.85	7.10	8.90	7.08	8.92
60	5.93	8.38	6.67	9.33	6.63	9.37
80	5.67	8.98	6.24	9.76	6.19	9.81
100	5.43	9.65	5.81	10.19	5.76	10.24
120	5.22	10.41	5.39	10.61	5.34	10.66
140	5.03	11.25	4.99	11.01	4.94	11.06
160	4.88	12.20	4.54	11.46	4.56	11.44
180	4.71	13.27	4.18	11.82	4.19	11.81
200	4.57	14.46	3.82	12.18	3.84	12.16
220	4.45	15.80	3.49	12.51	3.52	12.48
240	4.35	17.31	3.17	12.83	3.21	12.79
260	4.25	18.99	2.86	13.14	2.93	13.07
280	4.17	20.89	2.58	13.42	2.66	13.34
300	4.09	23.01	2.32	13.68	2.42	13.58
320	4.03	25.40	2.06	13.94	2.19	13.81
340	3.97	28.07	1.84	14.16	1.98	14.02
360	3.91	31.07	1.63	14.37	1.79	14.21
380	3.86	34.44	1.44	14.56	1.61	14.39
400	3.82	38.22	1.26	14.74	1.45	14.55
420	3.78	42.46	1.10	14.90	1.31	14.70
440	3.75	47.21	.94	15.06	1.18	14.82
460	3.72	52.55	.83	15.17	1.06	14.94
480	3.69	58.54	.72	15.28	.95	15.05
500	3.67	65.26	.61	15.39	.85	15.15
520	3.65	72.80	.53	15.47	.76	15.24
540	3.63	81.26	.45	15.55	.68	15.32
560	3.61	90.95	.38	15.62	.61	15.39
580	3.60	101.40	.32	15.68	.55	15.45
600	3.58	113.34	.27	15.73	.49	15.51
620	3.57	126.75	.22	15.78	.44	15.56
640	3.56	141.79	.19	15.81	.39	15.61
660	3.55	158.67	.16	15.84	.35	15.65
680	3.54	177.60	.13	15.87	.31	15.69
700	3.54	198.85	.11	15.89	.28	15.72
720	3.53	222.69	.08	15.92	.25	15.75
740	3.52	249.44	.06	15.94	.22	15.78
760	3.52	279.45	.06	15.94	.20	15.80
780	3.51	313.13	.05	15.95	.18	15.82
800	3.51	350.91	.03	15.97	.16	15.84

derivative helps to translate percentage expectancy differences into rating differences; in the latter it provides a shortcut for cumulative averaging. In this section we shall look at the derivative in relation to cumulative averaging, beginning with the simpler case of interval ratings. A game-by-game rating process will be used, both for its simplicity and because averaging the competition in a ratio system is problematic.

Cumulative averaging for a percentage score is illustrated by

$$\Delta P = \frac{S - P_e}{N + 1}, \quad (67)$$

where  $S$  represents the new game score. The process is recursive because the variables  $P_e$  and  $N$  will be used in the next step with different values. The subscript for the original percentage score indicates expectancy in some contexts. The advantage over ordinary arithmetic averaging is that we do not have to write out the entire result at each step, but we do have to keep track of  $N$  separately. A similar process can be applied to linear ratings. If the new rating is

$$R_n = ER_c + K(2P - 1), \quad (68)$$

then using the original rating and the new mean opposition rating, we can solve for  $P_e$  in

$$R_o = ER_c + K(2P_e - 1). \quad (69)$$

By subtraction

$$\Delta R = 2K(P - P_e), \quad (70)$$

and substituting for the change in percentage by (67)

$$\Delta R = \frac{2K(S - P_e)}{N + 1}. \quad (71)$$

We get the same result by applying the constant derivative of the basic formula,  $2K$ , to the change in percentage score (67). For linear ratings, then, the cumulative average form of the established rating is the same as its differential form.

The differential form of the established rating for ratio systems is easily determined. The derivative of the basic ratio formula (16) with respect to  $P$  is

$$R' = \frac{ER_c}{(1 - P)^2}. \quad (72)$$

Applying this to the change in percentage score (67) gives

$$\Delta R = \frac{ER_c(S - P_e)}{(1 - P)^2 \cdot (N + 1)} \quad (73)$$

as the differential. The analogous formula for cumulative averaging is somewhat more complicated. Given a new ratio rating

$$R_n = ER_c \frac{P}{1 - P} \quad (74)$$

we can calculate expected percentage scores from the old rating and the new mean opposition rating as

$$R_o = ER_c \frac{Pe}{1 - Pe}. \quad (75)$$

The rating change follows as

$$\Delta R = \frac{ER_c(P - Pe)}{(1 - P) \cdot (1 - Pe)}. \quad (76)$$

Substituting for the percentage change in the numerator by (67),

$$\Delta R = \frac{ER_c(S - Pe)}{(1 - P) \cdot (1 - Pe) \cdot (N + 1)}. \quad (77)$$

This is the change formula for cumulative averaging in a ratio system. We assume, for the sake of comparison with the differential, that the mean opposition rating does not change. Since the percentage score is close to its expected value for a large sample, especially when rating by single games, we conclude that the differential provides a reasonable approximation to cumulative averaging in a ratio system.

### 13 The Berkin System

The Berkin System is a novel rating system. It is based on the idea of weighted ratings, an idea introduced by Berkin in 1965.<sup>27</sup> The Berkin System was a candidate for official recognition by FIDE, but adoption of the Elo System in 1970 led to its neglect. We have taken some liberties in presenting the system, but the essential idea remains intact. Berkin actually spoke of weighted *points*, but since his points are weighted by ratings, it amounts to the same thing. The weighted average of opposition ratings is

$$\sum \frac{R_c S}{W}. \quad (78)$$

Replacing the opposition average in (16) and reducing gives the simple ratio formula

$$R = \sum \frac{R_c S}{L}, \quad L > 0. \quad (79)$$

Only losses are recorded in the denominator, with draws counting as half-point losses. The numerator in effect does not include the ratings of winning opposition ratings, which are weighted by zero, but it does include the ratings of drawing opponents, weighted by the usual half point. This absence of data for losses is compensated in the calculation of opposition ratings, without the duplication characteristic of other systems. As a consequence (79) can be calculated

---

<sup>27</sup>E1, 8.62

simultaneously for a completing field, and the Berkin is the only recognized ratio system that can claim that advantage. A case can furthermore be made for the system as an improvement over ordinary ratio systems. For a given percentage score  $P$ , if the opposition ratings are all the same, a Berkin rating yields the same result as a standard ratio rating. Consider now a group of opponents with ratings that vary, and let us suppose that the percentage score against these opponents is  $P$ , such that  $0 < P < 1$ . The standard ratio formula yields a single rating regardless of the individual outcomes, while the Berkin formula yields a variable rating depending on the scores against individual opponents. The latter will be higher than the standard result if the wins are against higher ratings, lower if the wins are against the lower ratings. The Berkin rating, in short, extracts more information from precisely the same data, and the conclusion seem inescapable that it is a better statistic.

The efficiency of the revised system is demonstrated as follows: Multiplying Formula (79) by  $L$ , and expressing  $L$  as the total of lost points,

$$\sum[R(1 - S)] = \sum[R_c S]. \quad (80)$$

Subtracting the right side,

$$\sum[R(1 - S)] - \sum[R_c S] = 0. \quad (81)$$

The difference of totals may be regarded as a total of differences for individual pairings,

$$\sum[R(1 - S) - R_c S] = 0, \quad (82)$$

which may also be expressed as a mean

$$E[R(1 - S) - R_c S] = 0. \quad (83)$$

Again we have a mean formula to which Gauss's principle can be applied. With zero dropping from the sum of squares,

$$\sum[R(1 - S) - R_c S]^2 \quad (84)$$

is a minimum over real  $R$  where  $R$  is calculated by (79). The weighted ratings may be said to be mutually predictive.

Avoiding the difficulties that arise from opposition averaging, we derive a change formula from (79) for single-game events. Scoring  $S$  against the new opposition rating  $R_{cn}$  gives the new rating,

$$R_n = \frac{\sum(R_c S) + R_{cn} S}{L + 1 - S}. \quad (85)$$

The original rating may be given by an expression equivalent to (79) with the same denominator as (85),

$$R_o = \frac{\sum(R_c S) + R_o(1 - S)}{L + 1 - S}. \quad (86)$$

Subtracting  $R_o$  from  $R_n$  gives

$$\Delta R = \frac{R_{cn}S - R_o(1 - S)}{L + 1 - S}, \quad L + 1 - S > 0. \quad (87)$$

The denominator in this expression increases only with a loss. It can be maintained at a constant value from one game to the next when it is large enough to avoid significant error, giving

$$\Delta R = \frac{R_{cn}S - R_o(1 - S)}{N_o + 1 - S}, F \quad (88)$$

which is a reasonable change formula. Using Elo's blending process, we can simply write

$$\Delta R = \frac{R_{cn}S - R_o(1 - S)}{N_o}, \quad (89)$$

and for multiple results we can substitute  $W$  for  $S$ , and  $L$  for  $(1 - S)$ . Now with slightly different notation we have Elo's formula (56), representing the established rating formula of the Berkin System.<sup>28</sup>

$$\Delta R = \frac{R_cW - RL}{N_o}. \quad (90)$$

## 14 A Progressive System

In a progressive rating system players never lose rating points. This would be an obvious boon to organized chess if it were not for questions of accuracy. A workable progressive system would require careful management at the very least. Virtually any rating system can be made progressive simply by ignoring negative results, but the Berkin System seems especially suitable for this adaptation. The formula

$$R = \frac{\sum(R_cS)}{N_o} \quad (91)$$

is equivalent to the basic Berkin formula (79) when  $L$  reaches the value of the arbitrary constant. As we saw in Section 8, the sample size in cumulative averaging can be maintained at a constant value, if sufficiently large, with negligible loss of accuracy. Thereafter the change formula for game-by-game results is

$$\Delta R = \frac{R_cS}{N_o}, \quad (92)$$

which is always nonnegative. Note that the basic performance formula (79) produces rating decreases for losses. This can be avoided by applying (92) from the outset of the rating process, which is justified by the fact that after  $N$

---

<sup>28</sup>E1, 8.62

successive applications, where  $N$  may be very large, the number of lost points eventually reaches  $N_o$ , and

$$\Delta R_1 + \Delta R_2 + \dots + \Delta R_N = \frac{\sum [R_c S]}{N_o}. \quad (93)$$

If the inaccuracies are deemed acceptable, Formula (92) becomes an all-purpose progressive formula. A drawback of progressive systems is that the growth of ratings is exponential, and ratings may consequently become huge over time. A logarithmic version of the system is possible, using an approximation from calculus. The limit

$$\lim \frac{\Delta(\ln R)}{\Delta R} = \frac{1}{R} \quad (94)$$

as  $\Delta R$  goes to zero defines the derivative of a natural logarithm by the delta method. For small values of  $\Delta R$  this gives the approximation

$$\Delta(\ln R) \approx \frac{\Delta R}{R}. \quad (95)$$

Substituting by (92) into this formula,

$$\Delta(\ln R) \approx \frac{R_c S}{R N_o}. \quad (96)$$

If the rating variables are assumed to be logarithmic,  $R$  may be substituted for  $(\ln R)$ ,  $e^R$  for  $R$ , and  $e^{R_c}$  for  $R_c$ , giving

$$\Delta R \approx \frac{S \cdot e^{R_c - R}}{N_o}. \quad (97)$$

Ratings may be initialized to zero. The change in rating for a result between two players new to the system would be  $S/N_o$ . A newcomer defeating a player rated  $\ln N_o$  would gain one point.

## 15 Tests

Tests of a rating system, such as those offered by Elo in his main work<sup>29</sup> tend not to be worth the paper they are written on, mainly because a rating is a statistic. The predictions they offer in this respect are self-fulfilling. By analogy, a test of arithmetic averaging to determine whether it yields central values would be pointless. Ratings do not predict changes in playing strength. Rather, they assume that the playing strength exhibited by past results will be exhibited in future results. Their predictions are an extrapolation of demonstrated playing strength. If the ratings tend toward a long-term limit, their predictions will by and large hold true.

The typical rating test, especially among those that have been employed by this author, uses sequential calculations to determine convergence toward

---

<sup>29</sup>E1, 2.6

assumed playing strength. It must be said that such a demonstration is more a measure of cumulative averaging than of rating validity. For rating systems in general the convergence is slow, and differences from one rating system to another are of doubtful significance. A more meaningful test would employ simultaneous calculations, and rating systems that can be adapted to this process, such as linear systems and the Berkin System, have a distinct advantage. In such a test there is no need to postulate “true” rating strengths. The measure in this case is how well percentage scores match rating differences or ratios, and Gauss’ principle assures us that the match cannot be improved.

What a rating test that uses sequential ratings is actually measuring is progress toward consistency. Consistent ratings are predictive in a statistical sense, that is, one can predict from rating relationships what results would have occurred, even though many may already be known. The rate at which sequential ratings tend toward consistency is a measure of their efficiency in one respect, namely, how well they accommodate the averaging process, but it is not a certain measure of their overall efficiency. An averaging process designed to reveal long-term limits may not be adequate to the task of detecting changes in rating strength, and sequential ratings are especially vulnerable to such changes. Anti-deflationary measures, such as Elo’s “feedback,” may help by boosting the efficiency of the averaging process, but they are at best stopgap measures.

## 16 Skeptical Conclusions

Nathan Divinsky in *The Chess Encyclopedia* calls the Elo System “a mathematically sound and universally accepted (1970) rating system for chess players.”<sup>30</sup> The year refers to the adoption of the Elo System by FIDE. Aside from a 1965 contribution by Elo to *The Journal of Gerontology*, there has been virtually no peer review of the system beyond the world of organized chess. One of the few external references is to be found in *The Mathematics of Games*, by J. D. Beasley.<sup>31</sup> Beasley offered this scathing footnote on the work of the late Professor Elo:

Chess enthusiasts may be surprised that the name of Elo has not figured more prominently in this discussion, since the Elo rating system has been in use internationally since 1970. However, Elo’s work as described in his book *The rating of chessplayers, past and present* (Batsford, 1978) is open to serious criticism. His statistical testing is unsatisfactory to the point of being meaningless; he calculates standard deviations without allowing for draws, he does not always appear to allow for the extent to which his test results have contributed to the ratings which they purport to be testing, and he fails to make the important distinction between proving a proposition true and merely failing to prove it false. In particu-

---

<sup>30</sup>D

<sup>31</sup>B, p. 61

lar, an analysis of 4795 games from Milwaukee Open tournaments, which he represents as demonstrating the normal distribution function to be the appropriate expectation function for chess, is actually no more than an incorrect analysis of the variation within his data. He appears not to realize that changes in the overall strength of a pool cannot be detected, and that his ‘deflation control’, which claims to stabilize the implied reference level, is a delusion. Administrators of other sports (for example tennis) currently publish only rankings. The limitations of these are obvious, but at least they do not encourage illusory comparisons between today’s champions and those of the past.

The proof of the pudding, it has been said, is the actual operation of a rating system, and the Elo System has been grinding out chess ratings for over four decades now with hardly a grumble from the rating pool. One is tempted to say that the system works despite its theory rather than because of it. The reputation of the Elo System rests largely on its supposed ability to predict chess outcomes. There is even the occasional inquiry as to whether the system can predict outcomes in sports such as basketball, football, golf and soccer. As this treatise has attempted to show, the predictive powers of the Elo System are not due to its application of probability theory, which in the final analysis must be characterized as a misapplication, but rather to principles of averaging which have hardly been articulated elsewhere.

The main weakness of the Elo System arises from the scientist’s habit of overvaluing the artifacts of his profession, in this particular instance, probability distributions. The system is not so much an attempt to apply statistical principles to the rating problem as an effort to shape mathematical intuitions to the judgments of abstract theory. As a case in point, Elo began his development of the established rating logically enough with cumulative averaging, but the argument then takes a detour into the expectancy curve. Again, great pains are taken with nonlinear probability functions, only to find that over their “most used” portions they behave much like a linear system. And again, the development of the logistic system begins with a principle that has been offered in this treatise as the basis of ratio systems but is cast in the Elo System by a dubious multiplication of odds as a probability distribution. Equally dubious are the attempts to explain the association of rating difference with percentage expectancy by the overlapping of hypothetical normal distributions. As an alternative to these tortured arguments, this treatise has postulated simple relations between rating differences or ratios and relative performance.

Probability theory, as it happens, does explain much of the success of the Elo System, but theory of a different sort than its author took for granted. The understanding of percentage scores with respect to rating differences or ratios as tending to long-term limits is quite absent from the system, though application of the frequency theory of probability seems natural enough. If Elo misapplied theory, he also made considerable use of mathematical intuition, which he otherwise disparaged. The result is a system that is a marked improvement over

those that preceded it, but a system that falls short of the scientific rigor that Elo envisioned for it. The lesson perhaps is that no one system is likely to be the last word in statistical precision. What, then, lies in the future for chess rating systems? There is, to be sure, no predicting the winds of change, but the fascination of chess ratings lies in their controversial nature and their capacity for inspiring new ideas. Let us hope that the pronouncements of experts do not prematurely put an end to the controversy.

## A References

- B** J. D. Beasley, *The Mathematics of Games*, Oxford University Press, 1989.
- D** N. J. Divinsky, *The Chess Encyclopedia*, Facts on File, New York, 1990.
- E1** A. E. Elo, *The Rating of Chessplayers, Past and Present*, Arco, New York, 1978.
- E2** - "The International Chess Federation Rating System,"  
*Chess*, Sutton-Coldfield, July, August, October, 1973.
- E3** - "F.I.D.E Information: The Rating System in National Applications"  
(A Supplement to the FIDE Presentation of 1970)
- G** M. Gardner, *The Colossal Book of Mathematics*, W. W. Norton, New York, 2001.
- GJ** M. R. Garey and D. S. Johnson, *Computers and Intractability:  
A Guide to the Theory of NP-Completeness*,  
Feeman, New York, 1967.
- HW** D. Hooper and K. Whyld, *The Oxford Companion to Chess*, 2nd Edition,  
Oxford University Press, 1992.
- J** R. C. Jones, "Evaluating Competitive Performance with Rankings and Ratings,"  
Master of Science Thesis, University of Rhode Island, 1994.
- V** R. v. Mises, *Probability, Statistics and Truth*, 2nd Revised English Edition,  
Dover Publications, New York, 1957.

## B Downloads

Supporting programs and documentation are available from *ftp.ratingtheory.com* by means of an FTP client, such as the free download at *www.coreftp.com*. Anonymous FTP is not supported. Enter the user name

downloads@ratingtheory.com

and the password “Back2Basics”. The folders presented there contain files relevant to the corresponding sections. The Visual Basic files (.vb) show program structures but will not run on their own.

**A Revised Elo System** contains a module for a Visual Basic console program used to generate the values of Table 10: “Rating Changes Produced by a Win for Various Rating Differences.”

**Consistency** contains live Microsoft Excel programs in two versions, one for Excel 2007 with extension *.xlsx* and a second for earlier releases with extension *.xls*. The Excel program *CrossTable2007* shows the convergence of linear ratings by iterative calculations. The Excel program *Champs2008* shows the convergence of Berkin ratings. The files *Blank* and *HalfTable* are for experiments with linear ratings. The latter provides automatic entry of opposition scores. The text file *ReadMe* gives instructions for operating the demonstration programs by the default manual calculations. The text file *Entering Data* gives hints for input.

**Ordinal Ratings** contains two modules for a Visual Basic console program. It compares the minimizing of violations by two algorithms: (1) sorting by directed rank difference and (2) brute force.

**Ordinal Ratings 2** contains the Excel program *Mannheim 1914*, which addresses the controversy over the performances of Alekhine vs. Vidmar in that tournament.

**Ordinal Ratings 3** contains a module for a Visual Basic console program that calculates sequential ordinal ratings for a hypothetical playing field and shows the resulting correlations, which are depicted in the graph “Correlation of Calculated Ranks with Predefined Ranks by Spearman’s Rank Correlation Coefficient.”

**Probability in Rating Systems** contains a VB module for generating the values depicted in the graph “Score Probabilities for Ten Games.”

**Sequential vs. Simultaneous Ratings** contains a VB module for generating the values of the graph “Effect of Half Rating Changes.”