

Back to Basics in Chess Ratings

Royal C. Jones, Jr. ©2010

March 31, 2010

Contents

1	Beyond the Elo System	2
2	Probability in Rating Systems	3
3	Ordinal Ratings	5
4	Interval Ratings	9
5	Ratio Ratings	10
6	Sequential vs. Simultaneous Ratings	12
7	Established Ratings	14
8	Attenuation	16
9	Percentage Expectancy	19
10	Consistency	21
11	Methods of Calculation	26
12	A Test	28
13	Skeptical Conclusions	29
A	The Berkin System	31
B	A Progressive System	32
C	Rating by Single Games	34
D	References	36
E	Downloads	37

1 Beyond the Elo System

Rating systems in their modern mathematical form first appeared in 1939 in a system used by the Correspondence Chess League of America. The influential Ingo System of West Germany followed in 1948, named by its originator Anton Hoesslinger (1875–1959) for his home town, Ingolstadt in Bavaria.¹ It establishes the basics of ratings in a remarkably simple formula,

$$R = ER_c - (Pct - 50), \quad (1)$$

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. Hoesslinger appears to have relied largely on intuition in developing his system, which manages nevertheless to be theoretically provocative. The actual development of rating theory took a different tack about 1960 with the introduction of probability formulas. The main proponent of this idea was Arpad E. Elo, one of the founders of the United States Chess Federation (USCF), whose system was subsequently adopted by the International Chess Federation (FIDE). The paradigm shift that brought the application of probability theory to the rating of chessplayers proved irresistible to mathematicians. About the same time that Elo was developing his system, similar ideas were afloat in Australia.²

Mathematicians should keep in mind, however, the essential nature of chess ratings. One is tempted to think of ratings as measurements of performance, in the same sense as measurements of physical phenomena. As a trained physicist Elo was especially prone to this interpretation. 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. Such speculations lead 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 existence of alternatives seems problematic by Elo's own standard. By analogy with scales of measurement, Elo distinguished three types of rating systems: *ordinal*, *interval*, and *ratio*. This classification is convenient enough for describing the different statistical methods that arise from rating theory and will be utilized in the following pages. But first the issue of probability will be revisited.

¹HW, "rating"

²E2, Part 1

2 Probability in Rating Systems

The basic 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.³ But what exactly 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.⁴

Performance in this light does not seem a very promising candidate for mathematical treatment. Fortunately, there is a simple definition of performance that is as old as the game. If a player outperforms the opponent, he/she wins and scores the point; if he/she loses, the opponent scores the point; and if a draw occurs, the point is divided. This simple definition need not lead to a simple-minded treatment. Under the heading “Sundry Theoretical Topics”⁵, 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 $P(\text{draw})$. 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 Formula (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”).

Although Elo presented Formula (2) almost as an afterthought, it provides a convincing demonstration of the “first and basic assumption” of his system.⁶ In the development of a rating theory, however, it is a dead end. Variation of performance thus described yields no useful definition of probability for rating theory, which is primarily concerned to relate percentage scores to ratings. The crucial percentage score is the mean value .65, which becomes increasingly prominent as sample size increases. The position taken here is that this long-term value is the link to probability theory.

Elo’s rather 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. Precisely how it does this is a crucial question for the system. We are shown two overlapping distributions⁷ and are told that the shaded portion of one “represents the

³E1, 1.31

⁴E1, 1.32

⁵E1, 8.9

⁶E2, Part 1, “Form Varies”

⁷E1, 8.23

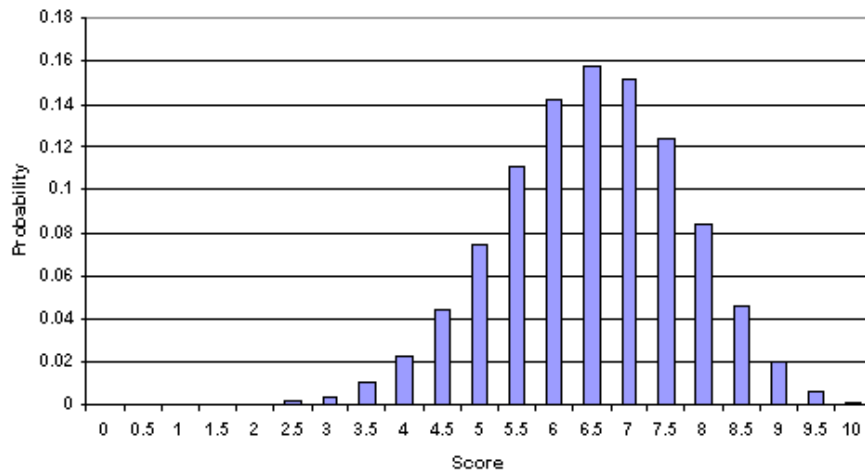


Figure 1: Score Probabilities for Ten Games

probability that the lower rated player will outperform the higher.” Apparently, if a player’s performance is greater than that of the opponent, he/she wins; if the performance is lower, he/she loses. This argument, aside from the objection that it leaves draws completely out of account, hangs on a tenuous concept of performance. Outperforming the opponent seems equivalent in every respect to winning; yet this would suggest a binomial distribution, if not trinomial.

For Elo the Percentage Expectancy Curve was patently a probability function, and he could make no sense of the objection that it was not. To avoid semantic arguments, the objection is better taken as a distinction of terms. 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. A more precise use of the term is restricted 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, applying Occam’s razor wherever needed. Regrettably, the bell curve has become an icon of chess ratings. What follows is a belated attempt to demonstrate a more reasonable theoretical basis.

3 Ordinal Ratings

Ratings that are based on an ordinal scale typically take the form of *ranks*, though other forms are admissible, e.g. percentiles. 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 to be NP-complete.⁸ That is, barring an earth-shaking breakthrough, there is simply no ranking algorithm to be discovered 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

$$rank(dominant) - rank(subordinate).$$

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 consists simply of sorting 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

⁸GJ

outdegree minus *indegree*.) The algorithm should be applied as a tiebreaker whenever possible, in which case the dominance relations among players with equal ranks are considered.

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

$$W_d(A) - L_d(A) < W_d(B) - L_d(B), \quad (3)$$

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 dominant or subordinate). The definition may be generalized to include the player himself, but this serves no useful purpose. 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 , which may be demonstrated as follows: 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 Formula (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 $W_d - L_d$ and P_d does not minimize violations but is a useful algorithm for reducing violations to a near minimum. A simulation by the present author pits this algorithm against a brute-force method (See “Downloads”). It produces a sequence of one hundred round-robin tournaments, each consisting of ten players. The players may be judged to be of equal strength inasmuch as every dominance relation is decided by the equivalent of a coin toss. For each tournament, first, the players were ranked by the dominance statistics developed here, using tiebreakers by the same method where possible, and the total number of violations produced by this ranking was noted as T1. Second, a ranking was found by a brute-force method that minimized violations, and the resulting number of violations was noted as T2. The upshot was an array of one hundred values for T1 and a corresponding array of one hundred values for T2. The mean value for T1 was 11.75, and for T2 it was 9.3. Considering the onerous nature of the brute-force method, there was not a great deal of improvement over the method based on dominance statistics. A method similar to the latter was shown by the author to be more efficient in reducing total violations, at least in random simulations, than more sophisticated probability algorithms.⁹

The principle of minimizing directed rank differences explains the efficiency of a round-robin tournament, which has long been regarded as a rational means for ordering the performances of its participants. The percentage scores resulting from a round robin, including half points for draws, are equivalent to dominance percentages, and their descending order consequently minimizes directed rank differences. The principle may 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.¹⁰ 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 Neustadt 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.

The dominance percentage statistic suggests a natural ordering of performances, expressed by the relation

$$Q \approx P_d, \tag{9}$$

where Q is the quantile (e.g., percentile) position of the particular value P_d in the competing field, which is almost never exact. The relation is analogous to the basic formulas of more advanced interval and ratio rating systems, to be introduced under subsequent headings. It motivates an algorithm in which quantile estimates are adjusted as outcomes occur. In this sequential ranking system, a list of opposition quantiles is maintained in ascending order for each player, say the quantiles of his/her last n opponents. Using the player’s point score against the current roster of opponents, a quantile rank based on Formula

⁹J, 2.2

¹⁰HW, “auxiliary scoring methods”

(9) may be estimated. Suppose a player has in the list of n opposition ranks the quantile values

$$Q_1, Q_2, Q_3, \dots, Q_{n-1}, Q_n$$

such that $Q_i \leq Q_{i+1}$ and that he/she has scored s points against the n opponents. The player's quantile would be calculated to fall above Q_s , but below Q_{s+1} , perhaps as the mean of these two values. The calculated quantile is then used in subsequent estimates of opposition quantiles, and the algorithm proceeds in self-corrective fashion.

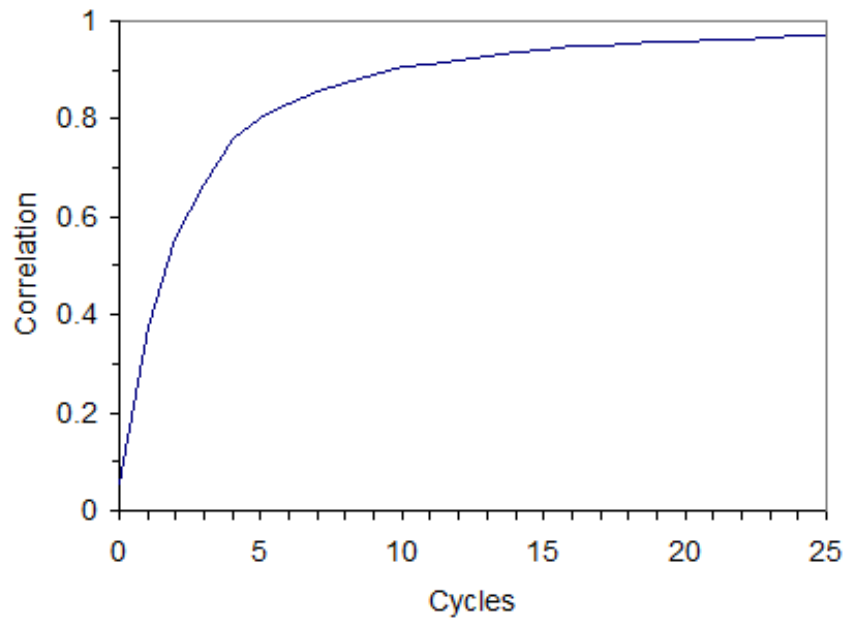


Figure 2: Correlation of Calculated Ranks with Predefined Ranks by Spearman's Rank Correlation Coefficient

This scheme is implemented in a program available by download from *ftp.RatingTheory.com* in the folder "Ordinal Ratings 2" (See "Downloads"). The program generates outcomes based on an arbitrary ranking of 400 players in 25 cycles of 200 games each, but users may change the experimental parameters. Results are calculated in the form of quantiles, which are then used to calculate rankings at the end of each cycle. The correlations between arbitrary rankings and calculated rankings are depicted in the graph above. It appears that a small list of opposition quantiles, three or four elements for each player instance, is adequate for generating plausible correlations. The output depicted is based on four elements in each list. Note that generating rankings from scratch presents subtle problems for the programmer. These are largely overcome by assigning initial quantiles as random values.

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 clearly $1 - P$, 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 the source of the Ingo System's power. We can generalize this discovery by a principle of interval systems which states that *differences in rating reflect differences in percentage score*. A rating formula that captures this principle is

$$R = ER_c + K(P - P_c). \quad (10)$$

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

$$2P - 1 = \frac{W - L}{N} \quad (11)$$

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

The effect of Formula (10) 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

$$R - ER_c = K(P - P_c). \quad (12)$$

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

$$E[R - R_c] = E[K(S - S_c)], \quad (13)$$

where the relative score, $S - S_c$, in chess evaluates to 1, -1, or 0 (for a win, loss, or draw respectively). 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. Proof of the efficiency of linear rating systems relies on the principle established by Gauss as the first step in his method of least squares: *For a given set of real values, the sum of squared deviations from a real variable is an absolute minimum where the variable is the arithmetic mean of the set of values*, which can be demonstrated by the mathematically trained as an exercise in differential calculus.

If the rating of Formula (10) is represented as the mean

$$R = E[R_c + K(S - S_c)], \quad (14)$$

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 (10) 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. This approximation is optimal for ratings calculated by the general linear formula, regardless of the consistency of data on which the ratings are based.

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 (10) is

$$R = ER_c \frac{P}{P_c}, \quad P_c > 0, \quad (15)$$

$$= ER_c \frac{P}{1 - P}, \quad P < 1, \quad (16)$$

$$= ER_c \frac{W}{L}, \quad L > 0, \quad (17)$$

with the same variables as the interval formula. Taking the reciprocal of this relation,

$$\frac{ER_c}{R} = \frac{P_c}{P}, \quad (18)$$

and from this we conclude by Gauss's principle that

$$\sum \left(\frac{R_c}{R} - \frac{P_c}{P} \right)^2, \quad R > 0, P > 0,$$

is a minimum over real values of R for the calculated value. Rating ratios may thus be said to predict relative performance, though in a somewhat weaker sense than in the case of interval ratings. Here relative performance cannot be defined for individual games because of the possibility of division by zero.

Elo was much impressed with the possibilities of a ratio scale, though perhaps for the wrong reasons. In the realm of physical measurements, ratio scales represent a considerable advance over interval scales. In the realm of statistics, their primary advantage is that probabilities can be expressed over a full range of rating ratios. In any case, Elo went on to develop a ratio version of his rating system which was eventually implemented by the USCF.

Elo's development of his ratio system was very different from that of his interval system, though he aimed in the end to show the near equivalence of the

two systems. As we shall see, his ratio system follows, roughly at least, from the basic ratio principle of Formula (15). We begin with the performance rating formula, which is oddly missing from the 1978 treatise but may be inferred from Formula (46):

$$P(D) = \frac{1}{1 + 10^{-D/2C}}. \quad (19)$$

Solving for D gives

$$D(P) = 2C \cdot \log_{10} \frac{P}{P_c}, \quad (20)$$

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

$$R = ER_c + 400 \cdot \log_{10} \frac{P}{P_c}, \quad P > 0, P_c > 0. \quad (21)$$

This formula follows directly from (15), assuming a geometric mean for the opposition ratings, by taking the logarithm of each side to a small base b . There is no special notation for ratings as logarithms. The base may be calculated from

$$\log_b R = 400 \cdot \log_{10} R. \quad (22)$$

By the rule of logs

$$\log_b R = \frac{\log_{10} R}{\log_{10} b}, \quad (23)$$

and setting equals to equal,

$$\frac{\log_{10} R}{\log_{10} b} = 400 \cdot \log_{10} R. \quad (24)$$

Consequently,

$$\log_{10} b = \frac{1}{400} \quad (25)$$

and

$$b = 10^{1/400} \quad (26)$$

or about 1.00577. Formula (20) as a logarithm of (15) assumes a geometric mean for the opposition ratings of Formula (15), but the mean proposed is arithmetic to take advantage of least squares approximation. This is a potential source of error in the Elo System. Possible improvements are discussed at “Rating by Single Games.” It is also worth noting that taking the logarithm of (15) does not produce a linear formula equivalent to (10) since relative performance in the result would be

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

Consequently, the efficiency of Formula (18) cannot be demonstrated from the theory developed for interval ratings. Another ratio system is the Berkin System, 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 considerably, 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, either by deliberately allowing one's rating to fall in order to qualify for prizes in lower-rated tournament sections, a practice known as *sandbagging*, or by selecting weaker opponents, such as may be met in the early rounds of a tournament, to assure a slow but steady rise in one's rating.
- While some ratings are more or less stable, others are rising rapidly, and interaction of the two sorts 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 may be unfortunate. 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 $\frac{1}{N}$, 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 generally 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.

A more rigorous application of the rating process involves simultaneous calculations for a defined data set. In 1969 such an application produced the first International Rating List.¹¹ Recursive calculations on a computer were applied to the complete interplay of 210 contestants over the previous three years. This was regarded primarily as a method for initializing the rating pool, but the effect was to produce a self-consistent set of ratings with clearly defined boundaries. Linear programming, as exemplified by this method, has the drawback of being

¹¹E2, Part 3

computationally intensive, and it is an open question whether the masses of data processed by a large rating system could be handled in this manner.

As a first step toward simultaneous calculations, it seems reasonable to deal with rating adjustments between pairs of contestants. Sequential ratings typically consist of rating adjustments based on expected performance against prevent opposition ratings. These adjustments are mirrored in the opposition ratings. Intuitively, we should be able to improve the accuracy of ratings by halving each rating adjustment. The simulation summarized in the graph below (See “Downloads”) suggests that this idea works in the initial stages of interplay, but that the advantage is eventually dissipated. The simulation consists of random interplay among a field of 100 players, which is repeated over sequences of various numbers of games: 100, 200, . . . , 800. The higher ranked player invariably wins, and results are rated sequentially by linear ratings. At the conclusion of each repetition an error statistic is calculated for all pairings in the field, namely, the root mean square of the deviation of actual relative performance, $W - L$, from expected relative performance based on the generated ratings.

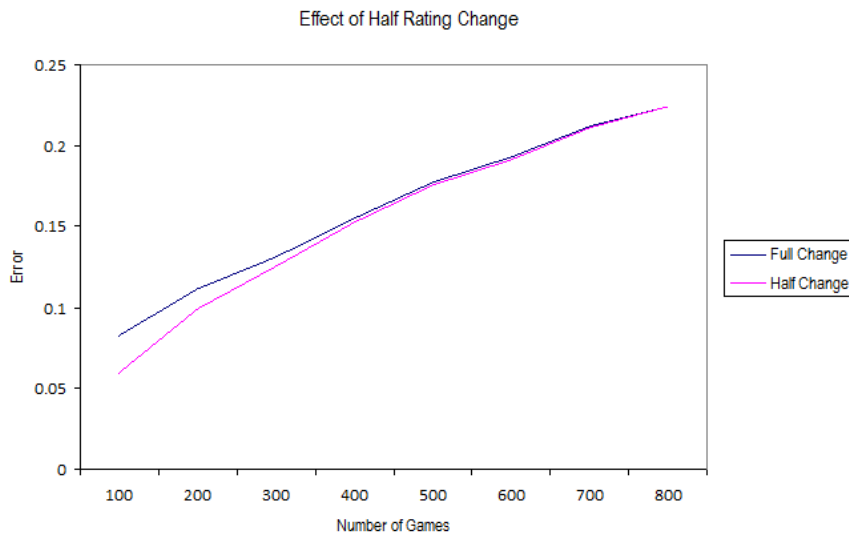


Figure 3: Effect of Half Rating Change

7 Established Ratings

Established rating formulas in the Elo System are thought to be suited for ratings based on reasonably large samples, where percentage expectancy can be reliably inferred from rating differences or ratios. This view is at odds with actual rating practice. Established formulas are developed as close approximations of performance formulas and require no particular assumptions about probability. They are generally easier to evaluate as change formulas, and therein lies their primary usefulness. The key concept in this conundrum is *percentage expectancy*, which will be explored in more detail further on. It may be taken for the moment as a ponderous term for a simple concept: namely, percentage score viewed as a function of rating difference or rating ratio. In the case of interval ratings, if the performance rating is a function of the mean opposition rating and the percentage score against that opposition, as

$$R = ER_c + K(2P - 1), \quad (27)$$

then the percentage expectancy for the difference $R - ER_c$ is

$$P_e = \frac{R - ER_c}{2K} + .5, \quad (28)$$

which is no more mysterious than algebraic manipulation.

The established formula is essentially a method for combining ratings. As we have seen, rating formulas may be written as arithmetic averages. If we combine a rating R_o based on N_o games with a rating R based on N games, the new rating may be written as the weighted average

$$R_n = \frac{R_o N_o + RN}{N_o + N}. \quad (29)$$

This formula can be applied in cumulative fashion, with the original sample N_o increasing to infinity. At some point in this process, if N_o is sufficiently large, it can be arbitrarily maintained at a constant value with a small loss of precision. The term *sampling weight* will be used hereafter for the constant N_o to distinguish it from an ordinary sample size. The formula then becomes

$$R_n = \frac{R_o(N_o - N) + RN}{N_o}, \quad (30)$$

which is Elo's "blending process".¹² Writing the original rating R_o as the identity

$$R_o = \frac{R_o(N_o - N) + R_o N}{N_o}, \quad (31)$$

a change formula follows immediately as

$$\Delta R = R_n - R_o = (R - R_o) \frac{N}{N_o}. \quad (32)$$

¹²E1, 8.63, described in 8.25

This formula simplifies the averaging process and restricts the sample weight of the original rating by attenuation. The simplification is pursued further by substituting linear formulas for R and R_o ,

$$R = ER_c + K(2P - 1), \quad (33)$$

and for the original rating in terms of percentage expectancy,

$$R_o = ER_c + K(2P_e - 1), \quad (34)$$

giving the simple result

$$\Delta R = 2K(P - P_e) \frac{N}{N_o}. \quad (35)$$

Results are more complicated with nonlinear formulas. Elo chose a differential version of Formula (32),

$$\Delta R = R'(P - P_e) \frac{N}{N_o}, \quad (36)$$

where R' is the derivative of the basic rating formula with respect to P . This simplifies to

$$\Delta R = \frac{R'(W - W_e)}{N_o}. \quad (37)$$

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

$$\Delta R = K(W - W_e), \quad (38)$$

which is Elo's established rating formula. The differential (37) is a reasonable approximation for small rating changes, which again requires a large value for the sampling weight. The derivative R' is the inverse of the derivative of the Percentage Expectancy Curve, which is approximated as the average slope of "the most used portion," roughly 1 percentage point over 8 points of rating difference.¹³ R' thus evaluates to the reciprocal of this approximation, 800 rating points. For logistic ratings, R' is the inverse of the derivative of Elo's logistic formula (46), which is his Verhulst formula (45). A precise value for R' is found by differentiating the logistic formula. As a function of D , this may be written

$$P = \frac{1}{1 + 10^{-D/2C}}. \quad (39)$$

It follows that

$$10^{D/2C} = \frac{P}{1 - P}. \quad (40)$$

Taking the common log of each side

$$\frac{D}{2C} = \log_{10} \frac{P}{1 - P}. \quad (41)$$

¹³E1, 8.25

Multiplying by $2C$ and converting to natural logs,

$$D = \frac{2C}{\ln 10} \cdot \ln \frac{P}{1-P}. \quad (42)$$

The derivative with respect to P follows as

$$\frac{dD}{dP} = \frac{2C}{\ln 10 \cdot P \cdot (1-P)}. \quad (43)$$

Since D is measured from a constant opposition rating, this is also the derivative of R with respect to P . Thus,

$$R' = \frac{2C}{\ln 10 \cdot P \cdot (1-P)}. \quad (44)$$

The inverse of (43) is

$$\frac{dP}{dD} = \frac{\ln 10 \cdot P \cdot (1-P)}{2C}, \quad (45)$$

which is a simplified version of Elo's Verhulst formula,¹⁴ although Elo's formula properly gives the derivative in terms of D .

The sampling weight in Elo's blending method is sometimes thought of as the maximum sample size of the original performance rating, which is to be balanced against the size N of a further performance sample for proper weighting of results. This analysis overlooks the recursive nature of the process, which is somewhat obscured by Elo's 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 $N = 5$.

8 Attenuation

Elo's blending method is an improvement on older systems that relied on performance samples of fixed size. The Harkness System took a rating average of the last four events,¹⁵ which allowed the unfortunate possibility of a winning result leading to a drop in rating. There are, however, improvements on Elo's

¹⁴E1, 8.43

¹⁵E1, 8.53

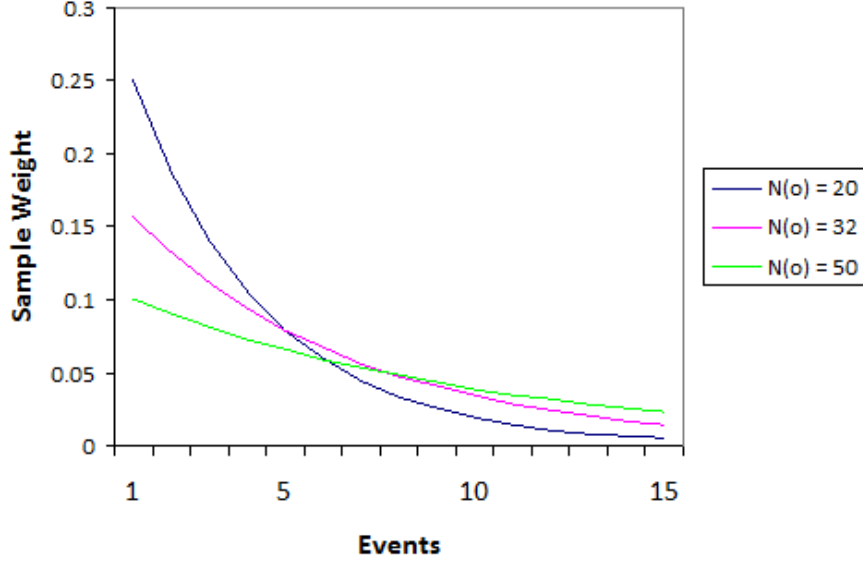


Figure 4: Attenuation of Sample Weight

improvement to be considered. Elo's method maintains a constant sampling weight at the expense of the original sample. Thus, in the blending method previously presented as

$$R_n = \frac{R_o(N_o - N) + RN}{N_o}, \quad (46)$$

the sample weight of the original rating,

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

clearly depends on the weight of the new sample. The size of the new sample must never exceed the size of the original. Somewhat better results obtain if sampling weight is allowed a temporary expansion at each step in the recursive process. The formula previously presented as

$$R_n = \frac{R_o N_o + RN}{N_o + N}, \quad (47)$$

serves to describe this process, except that here the sampling weight reverts to its constant value after each recursive step. This method of combining ratings more closely resembles the ordinary averaging process. The original sample weight is essentially independent of the new sample weight.

The primary benefit of a constant sampling weight, as we have seen, is attenuation of sample weights. In the Elo System this attenuation does not occur until the advent of established ratings, but logically there is no reason why it should not occur from the outset of the rating process. Attenuation can be achieved more directly through the application of an attenuating factor α , such that $0 < \alpha < 1$. At this point it is best to abandon Elo's notation in favor of one that is more mathematically descriptive. Our concern is to calculate a sequence of rating results based on cumulative averaging,

$$|R_1, |R_2, |R_3, \dots$$

The vertical line before each variable indicates cumulative results. The sequence is calculated using a corresponding sequence of variable sampling weights,

$$|N_1, |N_2, |N_3, \dots,$$

as well as a corresponding sequence of performance ratings for events $i = 1, 2, 3, \dots$

A formula that captures the cumulative nature of attenuated weighted sampling for sequential ratings would be

$$|R_i = \frac{\alpha |N_{i-1} \cdot |R_{i-1} + N_i \cdot R_i}{|N_i}, \quad (48)$$

where

$$|N_i = \alpha |N_{i-1} + N_i. \quad (49)$$

Analogous to Elo's notation, R_i is a performance rating based on N_i games. The variable $|N_i$ corresponds to Elo's constant N_o , and $|R_i$ to the new rating R_n . The process begins with the original performance rating, which is to say,

$$|R_0 = R_0 \quad (50)$$

and

$$|N_0 = N_0. \quad (51)$$

At each stage in the rating process, the sampling weight is decreased with multiplication by α and increased with addition of N_i . Thus,

$$|N_1 = \alpha N_0 + N_1, \quad (52)$$

$$|N_2 = \alpha^2 N_0 + \alpha N_1 + N_2, \quad (53)$$

and so forth, which is a geometric series. If N_i is assumed to be a constant N , then

$$\lim |N_i = \frac{N}{1 - \alpha} \quad (54)$$

as i goes to infinity. Interestingly, if we set α to

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

that is, the weight of the old rating in Elo's blending process (30), we find the limit to be N_o .

A change formula analogous to (32) is found by using the identity

$$|R_{i-1} = \frac{\alpha |N_{i-1} \cdot |R_{i-1} + N_i \cdot |R_{i-1}}{|N_i}. \quad (55)$$

It becomes clear that this formula is an identity by using the substituiton

$$N_i = |N_i - \alpha |N_{i-1}. \quad (56)$$

which follows from (49). The expanded equation quickly reduces to an identity.

Substracting the identity (55) from 48)

$$\Delta R = |R_i - |R_{i-1} = (R_i - |R_{i-1}) \frac{N_i}{|N_i}, \quad (57)$$

which is essentially a more subscripted version of (32), although here $|N_i$ is a variable rather than a constant.

We find, then, that a sampling constant is not necessary for the implementation of rating attenuation. Accurate attenuation by a factor of α can be maintained by Formula (57) from the very outset of the cumulative rating process. The downside is that sampling weight must be updated and stored at each step in the cumulative process, along with the cumulative rating itself.

9 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 $\frac{r}{n}$. This definition is at the heart of probability distributions, which do not fare well in the realm of chess ratings. The Percentage Expectancy Curve, patterned after the normal or Gaussian distribution or the similar logistic distribution, is not really a probability distribution at all. Instead, it arbitrarily assigns probability values to rating differences in imitation of these important distributions. The point is made clear by consideration of an actual distribution of pairings in a competing field, ordered by their algebraic difference. Each pairing is included twice: first, for the difference from the first player's standpoint; and second, for the difference from the second player's standpoint, making a symmetrical distribution. Can percentage expectancies be deduced from the parameters of the distribution? We could at best come up with probabilities for rating differences falling above or below specific values, which would tell us nothing about the relative performance associated with specific differences.

The second definition of probability is more promising in this context. This definition 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

$\frac{r}{n}$. The percentage scores encountered in rating systems may be regarded as estimates of this probability. 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.

Probabilities in a rating system are relative to pairs of contestants in the competing field. Rating systems generally assign ratings on the basis of relative performance as an estimate of these probabilities, either as differences in percentage score or ratios of percentage score. We may speculate that unease with the derivation of interval ratings from the normal curve led Elo to his logistic system, which takes a completely different approach. He began with the premise that the odds of player x to score over player z are

$$\frac{P_{xy}}{P_{yx}} \cdot \frac{P_{yz}}{P_{zy}} = \frac{P_{xz}}{P_{zx}}$$

where P_{xy} is the probability of x scoring over y , etc.¹⁶ This is based on the clearly false notion that probabilities are transitive. It is a commonplace observation in chess and other games that results are not transitive. If x defeats y , and y defeats z , it does not follow that x defeats z , even though the latter result may be in some sense expected. A similar observation holds for probabilities. Martin Gardner in one of his mathematical sketches¹⁷ reports on a set of four dice cleverly constructed to demonstrate this. In a game using these dice, a player selects a die with the idea of maximizing his chances. The second player is then able to select one of the remaining dice such that his odds of winning any roll-off against the first player are 2:1 in his favor. This is because the probabilities involved are not transitive. There is no “best” die among the four.

It could be argued that Elo was postulating the odds that would hold if probabilities *were* transitive. The point, in any case, is largely moot in view of the fact that his logistic system can be derived from the basic ratio formula (15). The inverse of this formula is percentage expectancy,

$$P_e = \frac{R}{R + ER_c}. \quad (58)$$

In a logarithmic system

$$P_e = \frac{b^R}{b^R + b^{R_c}} \quad (59)$$

for its base b . In Elo’s logistic system, as we saw in “Ratio Systems,”

$$b = 10^{1/400}. \quad (60)$$

Consequently,

$$P_e = \frac{10^{R/400}}{10^{R/400} + 10^{R_c/400}}. \quad (61)$$

¹⁶E1, 8.33

¹⁷G, “Nontransitive Dice and Other Paradoxes”

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

$$P_e = \frac{1}{1 + 10^{(R_c - R)/400}}. \quad (62)$$

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

$$P_e = \frac{1}{1 + 10^{-D/2C}}, \quad (63)$$

which is the logistic formula for Elo's Percentage Expectancy Curve.¹⁸

It was previously remarked that the advantages of a ratio rating system over an interval system are not pronounced in a statistical setting that does not involve physical measurements. A ratio system nevertheless has its advantages, most notably in the fact that its rating scale is unlimited. 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]" need not be taken seriously. There is a tendency, as Elo himself noted, for averaging to counteract the effect by the Central Limit Theorem.¹⁹ His speculation does suggest the interesting possibility of replacing ratings with quantiles (e.g., percentiles) in the basic linear formulas. Quantiles by their very nature are uniformly distributed in a random population, which would allow undistorted averaging. Manipulation of quantiles by linear formulas would act directly on the parameters of their distribution, although a continuous updating of their values would then be necessary.

10 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.²⁰ For the first International Rating List (see "Sequential vs.

¹⁸E, 8.43

¹⁹E1, 8.57

²⁰E1, 1.5-6

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.²¹ 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 following hypothetical tournament:

Table 1: Single Round Robin for Four Players

Players	A	B	C	D	wins	pct.
A	x	1	1	0	2	.667
B	0	x	1	1	2	.667
C	0	0	x	1/2	.5	.167
D	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).

Table 2: Matrix Representation of Table 1: Formula (10) ($K = 1$)

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

Table 3: Row Canonical Form for Table 2

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

Systems of ratio formulas, including the Berkin formula (88), are homogeneous, allowing only the zero solution (0, 0, 0, 0). For basic ratio ratings (15) there is a strategy for finding nonzero solutions where there is an undefeated player, whose rating is ordinarily undefined because of division by zero. Consider, for example, the addition of Player E to the above round robin with a

²¹E2, Part 3

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) as well as the notable advantage that the solution set can be assigned an arbitrary mean (here .25). Finally, there is the matrix representation of the system of Elo formulas, calculated by Formula (21) (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 (15)

	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 (88)

	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
Σ	1	1	1	1	1

These results are confirmed by iterative calculations on the round robin 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 (21)

	A	B	C	D	
A	3	-1	-1	-1	361.236
B	-1	3	-1	-1	361.236
C	-1	-1	3	-1	-838.764
D	-1	-1	-1	3	0

Table 8: Row Canonical Form for Table 7

	A	B	C	D	
A	3	-1	-1	-1	361.236
B	0	8	-4	-4	1444.944
C	0	0	48	-48	-11460.672
D	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 recently 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

$$R_f = \frac{R - R_{min}}{R_{max} - R_{min}}. \quad (64)$$

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

$$R_E = Elo_{min} + R_f(Elo_{max} - Elo_{min}). \quad (65)$$

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)

Linear Formula (10), 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

Berkin Formula (88), 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

Elo Formula (20), 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

Ratio Formula (15), 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.

11 Methods of Calculation

Percentage expectancy in the Elo System is associated with a probability distribution function. Rating changes are calculated with the derivative of this function. If it is true, however, that probabilities are to be viewed as long-term percentage scores, then every percentage score is associated with a probability, and rating changes may be calculated by ordinary averaging. We begin with a percentage score, such as might result from any chess event:

$$P_o = W_o/N_o, \quad (66)$$

which in chess has the special meaning of W_o points scored (including draws) in N_o games. The subscripts indicate original values since we are considering the process by which percentage score changes. For the next event, in which W points are scored out of N games,

$$P_n = \frac{W_o + W}{N_o + N}, \quad (67)$$

which is the new percentage score by cumulative averaging. Cumulative averaging can be applied recursively, which avoids having to write out the entire average at each step. The original percentage score (66) as an estimate of long-term percentage score was the expected score, which may be written in the equivalent form,

$$P_e = \frac{W_o + NP_o}{N_o + N}. \quad (68)$$

The change in percentage score now follows as

$$\Delta P = P_n - P_e = \frac{W - NP_o}{N_o + N}. \quad (69)$$

This can be converted to a rating change by using the constant derivative of the basic linear formula with respect to P , which is $2K$:

$$\Delta R = \frac{2K(W - NP_o)}{N_o + N}. \quad (70)$$

A change formula can also be derived from the basic linear formula (10). The new rating, based on a new percentage score, is

$$R_n = ER_c + K(2P_n - 1). \quad (71)$$

The original rating may be written using a percentage expectancy for the same mean opposition rating

$$R_o = ER_c + K(2P_e - 1). \quad (72)$$

Subtracting R_o from R_n ,

$$\Delta R = 2K(P_n - P_e). \quad (73)$$

We can substitute into Formula (73) using the percentage difference (69), giving the same result as in (70). There is no need to use a derivative.

Similar observations apply to ratio ratings. The derivative of the basic ratio formula (15) with respect to P is

$$R' = \frac{(R + ER_c)^2}{R_c}, \quad (74)$$

which can be used to convert (69) into a change formula for ratio ratings:

$$\Delta R = \frac{R'(W - NP_o)}{N_o + N}. \quad (75)$$

The derivative is not necessary for deriving ratio change formulas, as will be seen in a subsequent topic (“Rating by Single Games”). The formula developed there is based on single-game events:

$$\Delta R = \frac{[S(R + R_c) - R] \cdot (R + R_c)}{(1 - S)(R + R_c) + N_o R_c}.$$

Since it does not use the differential form of (75) it is likely to be more accurate, but it is interesting to compare the two formulas. First, applying (75) to single-game events,

$$\Delta R = \frac{R'(S - P_o)}{N_o + 1}, \quad (76)$$

We can write the write the expression $S - P_o$ in terms of percentage expectancy as

$$\frac{S(R_o + R_c) - R_o}{R_o + R_c}.$$

Substituting this expression back into Formula (76), along with the derivative for a single game by (74), and reducing,

$$\Delta R = \frac{R_o + R_c}{R_c} \cdot \frac{S(R_o + R_c) - R_o}{N_o + 1}. \quad (77)$$

Multiplying through and rearranging terms,

$$\Delta R = \frac{[S(R_o + R_c) - R_o] \cdot (R_o + R_c)}{R_c + N_o R_c}, \quad (78)$$

which is proportional to (114), the result obtained without a derivative.

12 A Test

Tests of a rating system, such as those offered by Elo in his main work²², tend not to be worth the paper they are written on, mainly because a rating is a measure of playing strength only in a metaphorical sense. Ratings are statistics, and the predictions they offer in this respect are self-fulfilling. A test of arithmetic averaging, by analogy, to determine whether it yields central values would be pointless. Ratings do not predict changes in playing strength: whether, for instance, a particular twelve-year-old will remain a novice all his/her life or become the next Bobbie Fischer. Ratings assume, rather, that the playing strength exhibited by past results will be the playing strength exhibited by future results. Their predictions are nothing more than an extrapolation of demonstrated playing strength.

Rather than pursue the topic announced in the title, it may be more profitable to look at rating statistics in the abstract. A less invidious simulation of the various rating systems is taken up under "Rating by Single Games." Let us designate rating differences or ratios as δ . Then for any rating system

$$P(\delta) = f(\delta), \tag{79}$$

which is to say, percentage expectancy is a function of rating difference or ratio. Not just any function will do if efficiency is a consideration, but we will keep our discussion general. There is also the inverse function,

$$\delta = f^{-1}[P(\delta)]. \tag{80}$$

Ratings are determined from particular instances of δ , which in turn are determined from actual percentage scores.

$$\delta_a = f^{-1}(P). \tag{81}$$

For an interval scale

$$R = R_c + \delta_a, \tag{82}$$

and for a ratio scale

$$R = R_c \cdot \delta_a. \tag{83}$$

Once a rating is determined, it may be viewed in relation to another instance of δ , call it δ_b . The percentage expectancy associated with this new instance is determined from Formula (79) as

$$P(\delta_b) = f(\delta_b). \tag{84}$$

A rating change can be determined from the change in δ_b , which in turn is determined from the change in $P(\delta_b)$. From Formula (80) it follows that

$$\delta_b + \Delta\delta_b = f^{-1}[P(\delta_b) + \Delta P(\delta_b)] \tag{85}$$

²²E1, 2.6

and

$$\Delta R = \Delta \delta_b. \tag{86}$$

These formulas apply to rating systems in general, though a rating system may not require all of them. A rating system can manage well enough using only (81) through (83), which express the general idea of a performance formula, thus forsaking the notion of percentage expectancy. Especially noteworthy is the argument of the inverse function in (85). This new percentage expectancy comes about as a consequence of a new result. Recalling the cumulative form of arithmetic averaging, Formula (69) in “Methods of Calculation,”

$$\Delta P = P_n - P_e = \frac{W - NP_o}{N_o + N},$$

where the new result is W points out of N games. This formula in its original context serves as a description of cumulative averaging. We have made the additional assumption that percentage scores in a rating system tend toward a limit. The original percentage score in this light is an estimate of probability and is consequently written as the percentage expectancy P_e . The new percentage score P_n is accordingly a revised estimate of probability.

13 Skeptical Conclusions

Alfred North Whitehead in his 1911 *Introduction to Mathematics* observed,

It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them. . . .

From a slightly different perspective, this is precisely a formula for the stagnation of civilization, for someone at some time must do the hard thinking upon which such important operations are built, especially if civilization is to advance by the overthrow of some thoughtless dogma or mindless ritual.

Nathan Divinsky in *The Chess Encyclopedia*²³ calls the Elo System “a mathematically sound and universally accepted (1970) rating system for chess players.” 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.²⁴ Beasley offers this scathing footnote on the work of the late Professor Elo:

²³D

²⁴B, p.61

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 particular, 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, on the other hand, 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.

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. If Elo misapplied theory, he also made considerable use of mathematical intuition, which in other contexts he 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. If there is any lesson to be learned from his celebrated work, it is that no single system is likely to satisfy the requirements of statistical precision. Rating systems in the past, as Elo notes, "received acceptance because they produced ranking lists which agreed generally with the personal estimates of rankings made by knowledgeable chess players."²⁵ Even now popular taste may have a role in deciding which system is to be

²⁵E2, Part 1

sanctioned by organized chess and how it is to be administered.

The principles of rating theory undoubtedly have applications beyond chess. As Elo said of his own system, it is “applicable to any type of competitive activity in which individuals or teams engage in pairwise competition.”²⁶ To this may be added applications for noncompetitive pairwise comparisons, as in opinion sampling for marketing research. One would hope that the current controversy is resolved before such wholesale applications. For some, however, the allure of rating theory lies in the controversy itself. It is a controversy that has not yet been played out in organized chess and a cautionary tale for all involved.

A The Berkin System

An interesting twist on ratio ratings involves the used of weighted ratings, an idea introduced by Berkin in 1965.²⁷ The Berkin System was a candidate for official recognition by FIDE, but adoption of the Elo System in 1970 led to its neglect. 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

$$\sum \frac{R_c S}{W}. \quad (87)$$

Replacing the opposition average in (15) gives the simple ratio formula

$$R = \sum \frac{R_c S}{L}, \quad L > 0. \quad (88)$$

Only losses are recorded in the denominator. The numerator in effect does not include the ratings of winning opposition ratings, which are weighted by zero. This loss of data is compensated in the calculation of opposition ratings, without the duplication of effect characteristic of other systems. The efficiency of the revised system is demonstrated as follows:

Multiplying Formula (88) by L , and expressing L as the total of lost points,

$$\sum [R(1 - S)] = \sum [R_c S]. \quad (89)$$

Subtracting the right side,

$$\sum [R(1 - S)] - \sum [R_c S] = 0. \quad (90)$$

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

$$\sum [R(1 - S) - R_c S] = 0, \quad (91)$$

²⁶E1, preface

²⁷E1, 8.62

which may also be expressed as a mean

$$E[R(1 - S) - R_c S] = 0. \tag{92}$$

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

$$\sum [R(1 - S) - R_c S]^2 \tag{93}$$

is a minimum over real R where R is calculated by (88). The weighted ratings may be said to be mutually predictive.

Anticipating the next section, it will be useful to derive a change formula from (88) for single-games events. Scoring S against the new opposition rating R_{cn} gives the new rating,

$$R_n = \frac{\sum(R_c S) + R_{cn} S}{L + 1 - S}. \tag{94}$$

The original rating may be given by an expression equivalent to (88) with the same denominator as (94),

$$R_o = \frac{\sum(R_c S) + R_o(1 - S)}{L + 1 - S}. \tag{95}$$

Subtracting R_o from R_n gives

$$\Delta R = \frac{R_{cn} S - R_o(1 - S)}{L + 1 - S}, \quad L + 1 - S > 0. \tag{96}$$

The denominator in this expression increases only with a loss. It can be maintained at a constant N_o , as in the Elo System, though with somewhat less justification. The resulting formula,

$$\Delta R = \frac{R_{cn} S - R_o(1 - S)}{N_o}, \tag{97}$$

is simulated in Table 10. It is essentially the same formula cited by Elo.²⁸ See "Rating by Single Games" for a discussion of the table.

B 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 the problem of accuracy. A workable system would require careful management over long periods of time since a rating would progress over the entire career of a chess player. Virtually any rating system can be made progressive simply by ignoring negative

²⁸E1, 8.62

results, but the Berkin System seems especially suitable for this adaptation. The formula

$$R = \frac{\sum[R_c S]}{N_o} \quad (98)$$

is equivalent to the basic Berkin formula (88) when L reaches the value of the arbitrary constant. Thereafter the change formula for game-by-game results is simply

$$\Delta R = \frac{R_c S}{N_o}, \quad (99)$$

which is always nonnegative. Note that (88) results in rating decreases for losses. This can be avoided by applying (99) from the outset of the rating process, which is justified by the fact that after N successive applications, where N may be very large, the number of lost points eventually reaches N_o , and

$$\Delta R_1 + \Delta R_2 + \dots + \Delta R_N = \frac{\sum[R_c S]}{N_o}. \quad (100)$$

Formula (99) thus turns out to be an all-purpose progressive formula of remarkable simplicity. It is simulated in Table 10, which is discussed in "Rating by Single Games."

A drawback of progressive systems is that the growth of ratings is unconstrained, and ratings may consequently become huge over time. A rating by (99) increases exponentially, and a mere thousand wins or so against equal-rated opponents would bring it to the point of overflow as a double precision variable in ordinary computer languages. A logarithmic version of the system is possible, however, using an approximation from calculus. The limit

$$\frac{\Delta \log_a R}{\Delta R} = \frac{1}{R \cdot \ln a} \quad (101)$$

as ΔR goes to zero is a consequence of the derivative of a logarithmic function. For small values of ΔR then

$$\Delta \log_a R \approx \frac{\Delta R}{R \cdot \ln a}. \quad (102)$$

Substituting (99) into this formula,

$$\Delta \log_a R \approx \frac{R_c S}{R N_o \cdot \ln a}. \quad (103)$$

Even now calculations might require the use of unwieldy values. If the rating variables are assumed to be logarithmic,

$$\Delta R \approx \frac{S \cdot a^{R_c - R}}{N_o \cdot \ln a}, \quad (104)$$

which gives a manageable progressive formula for logarithmic ratings to base a .

C Rating by Single Games

The nonlinear character of ratio systems poses problems for averaging. It is not hard to come up with examples of inconsistency in the application of ratio formulas. We begin with an expression for percentage expectancy of a player rated R against an average opposition rating of ER_c , which follows algebraically from (15) as

$$P(ER_c) = \frac{R}{R + ER_c}. \quad (105)$$

The formula also applies to the individual opposition ratings, which may be written

$$P(R_c) = \frac{R}{R + R_c}. \quad (106)$$

Is it true for the N opponents of 105 that

$$N \cdot P(ER_c) = \sum [P(R_c)] ? \quad (107)$$

Generally, no. Perhaps this is the fault of the method of averaging we have chosen for (105), but the inaccuracy arises for any of the common means: arithmetic, geometric, or harmonic. It would seem preferable to avoid averaging altogether in the application of ratio ratings, and this is a strategy that has been followed by practitioners of the Elo System. A simple improvement is to calculate $P(D)$ by (19) for each of the opposition ratings, taking the sum of the results as W_e in the standard formula

$$\Delta R = K(W - W_e). \quad (108)$$

A more general solution for ratio ratings of single-game events begins with the description of cumulative averaging given in “Methods of Calculation.” Adapting (69) to a single result,

$$\Delta P_e = \frac{S - P_e}{N_o + 1} \quad (109)$$

which describes the incorporation of a score S into the running average. The original sample size here should not be confused with Elo’s constant. The original percentage expectancy refers in this case to the pregame average, but it might with equal plausibility refer to the percentage expectancy against a new opponent.

Another expression for change in percentage expectancy begins with the percentage expectancy of the original rating against a new opponent.

$$P_e = \frac{R}{R + R_c}, \quad (110)$$

which yields the equivalent expression

$$= \frac{R + \Delta R P_e}{R + R_c + \Delta R}$$

The new percentage expectancy resulting from a change in R would be

$$P_{ne} = \frac{R + \Delta R}{R + R_c + \Delta R}. \quad (111)$$

Subtracting (110) from (111) gives the new expression

$$\Delta P_e = \frac{(1 - P_e)\Delta R}{R + R_c + \Delta R}. \quad (112)$$

We now have two formulas for change in percentage expectancy, (109) and (112). Setting the two right sides equal,

$$\frac{S - P_e}{N_o + 1} = \frac{(1 - P_e)\Delta R}{R + R_c + \Delta R}. \quad (113)$$

We can now solve for change R as

$$\Delta R = \frac{[S(R + R_c) - R] \cdot (R + R_c)}{(1 - S)(R + R_c) + N_o R_c}. \quad (114)$$

If N_o in (109) is maintained as a constant instead of being incremented recursively, the solution becomes

$$\Delta R = \frac{[S(R + R_c) - R] \cdot (R + R_c)}{(1 - S)(R + R_c) + (N_o - 1)R_c}. \quad (115)$$

This is the ratio analogue to Elo's established formula applied to single games. The sum of changes in a multiple-game event would be applied to R_o as the pre-event rating. We can get some idea of the practical value of this formula by a simulation (See "Downloads") using random pairings among 800 players. The predefined outcome S for each pairing is compared with the expected score P_e for the ratings calculated by each of the simulated systems. Games are played in sequences of 1600. After each sequence the root mean square of $S - P_e$ is computed over the 1600 results, as shown in Table 10 and its corresponding chart. The predefined outcomes in the sequences of Table 10 are always a win for the higher-ranked player, but the user may substitute probability functions by uncommenting the appropriate lines in the playing field constructor. The constants for number of players, number of games played, etc., may also be changed.

As was forcefully argued in "A Test", one must be cautious about the conclusions drawn from such a simulation. Precise simultaneous calculations, it will be recalled, can be made for the Berkin System and for linear systems. For these systems at least, the differences observed in Table 10 seem to be an artifact of sequential calculations. Such differences can be overcome by a feedback process, such as that proposed by Elo.²⁹ Elo intended his feedback for rating what he called "exceptional performers," those which arise from changes in the

²⁹E1, 3.75

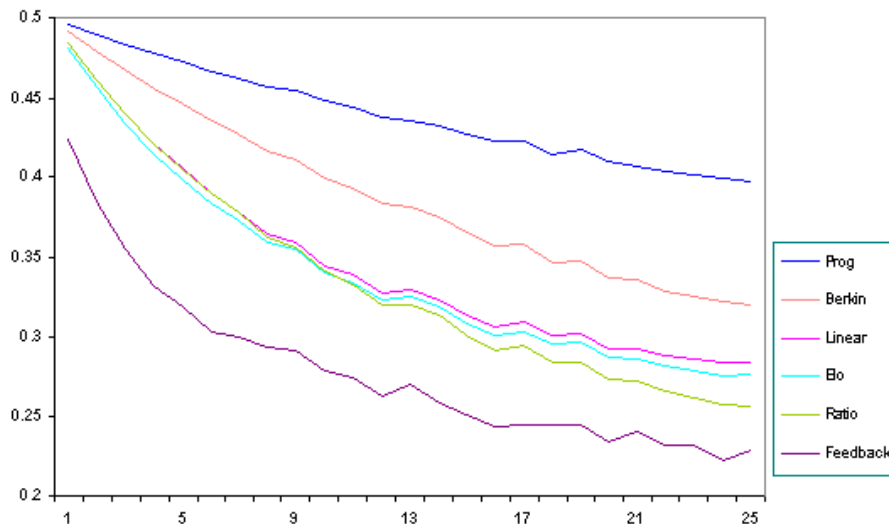


Figure 5: Graph of Table 10

underlying playing strength of rated players, but the process works equally well in the present context. The simulation provided as a download has a feedback loop in the subroutine RateBerkin. The feedback results of Table 10 can be obtained by uncommenting this loop. Exceptional performances are defined here as producing an absolute difference greater than .4 between expected and actual scores. As with other constants in the simulation, this may be changed by the user for experimental purposes.

D 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.
- 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.

E Downloads

Supporting programs and documentation are available from *ftp.ratingtheory.com* by means of an FTP client (If a basic FTP program is sufficient to one's needs, there is a free download at *www.coreftp.com*). Anonymous FTP is not supported. Enter the username *downloads@ratingtheory.com* and the password *Back2Basics*. The folders presented there contain files relevant to the corresponding sections:

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 spreadsheet *CrossTable2007* shows the convergence of linear ratings by iterative calculations. The spreadsheet *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 a Visual Basic console program consisting of two modules. It compares the minimization of violations by two algorithms: (1) sorting by directed rank difference and (2) brute force.

Ordinal Ratings 2 contains a Visual Basic console program that simulates the calculation of 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 Visual Basic console program for generating the values depicted in the graph "Score Probabilities for Ten Games."

Rating by Single Games contains a Visual Basic simulation of five systems. The values generated are shown in Table 10 with its accompanying graph.

Sequential vs. Simultaneous Ratings contains a Visual Basic console program for generating the values of "Effect of Half Rating Changes."

Table 10:
 Root Mean Square of Deviations
 of Actual Scores from Expected Scores
 in Simulated Random Pairings

	Progressive	Berkin	Linear	Elo	Ratio	Feedback
1.	0.49536	0.49153	0.48356	0.48132	0.48355	0.42399
2.	0.48946	0.47969	0.46164	0.45703	0.46164	0.38500
3.	0.48320	0.46726	0.43964	0.43339	0.43963	0.35499
4.	0.47790	0.45620	0.42121	0.41428	0.42102	0.33171
5.	0.47220	0.44601	0.40537	0.39830	0.40494	0.31896
6.	0.46668	0.43556	0.38998	0.38342	0.38940	0.30308
7.	0.46203	0.42713	0.37881	0.37285	0.37759	0.29972
8.	0.45721	0.41620	0.36469	0.35923	0.36267	0.29314
9.	0.45480	0.41091	0.35931	0.35413	0.35638	0.29121
10.	0.44768	0.39939	0.34501	0.34022	0.34078	0.27845
11.	0.44412	0.39324	0.33887	0.33417	0.33333	0.27480
12.	0.43844	0.38355	0.32765	0.32277	0.32036	0.26269
13.	0.43547	0.38115	0.32909	0.32441	0.32017	0.26992
14.	0.43275	0.37499	0.32313	0.31822	0.31285	0.25863
15.	0.42682	0.36523	0.31353	0.30806	0.30090	0.25078
16.	0.42275	0.35700	0.30623	0.30082	0.29171	0.24273
17.	0.42253	0.35817	0.30960	0.30368	0.29449	0.24517
18.	0.41468	0.34665	0.30023	0.29483	0.28350	0.24453
19.	0.41715	0.34716	0.30254	0.29611	0.28376	0.24493
20.	0.40986	0.33736	0.29279	0.28702	0.27365	0.23380
21.	0.40646	0.33521	0.29272	0.28630	0.27171	0.24062
22.	0.40434	0.32869	0.28794	0.28144	0.26584	0.23130
23.	0.40181	0.32509	0.28614	0.27897	0.26156	0.23268
24.	0.39903	0.32226	0.28320	0.27563	0.25750	0.22226
25.	0.39727	0.31956	0.28380	0.27601	0.25676	0.22866