The Legacy of Arpad Elo

The Development of a Chess-Rating System

Universiteit van Amsterdam
Faculteit der Psychologie
VRT-2
1 December 1998
Auteur: Sandra de Blécourt
Studentnummer: 9612858
Begeleiders: Han van der Maas, Eric-Jan Wagenmakers

Abstract

The Elo-rating system (Elo, 1978) is one of the largest applications of paired-comparison scaling. Every active chess player is assigned an Elo-rating indicating his or her chess proficiency. These ratings are used not only to make a ranking of all active players but also serve as a basis for selection of players for training and tournaments. Chess federations worldwide found ways to overcome difficulties that normally restrict the use of paired-comparison techniques in science (Batchelder & Bershad, 1979). Therefore the Elo-rating system yields interesting implications for psychology.

In this paper I will give an overview of the development of the Elo-rating system from its origin in the Thurstone Case V Model (Thurstone, 1927/1994) to recent investigations concerning the reliability of Elo-ratings under influence of factors such as age, routine, and playing style.

Contents

The Legacy of Arpad Elo

The Development of a Chess-Rating System

1. Introduction

The technique of paired-comparisons constitutes one of the main methodologies in quantitative psychology. Interestingly enough one of the largest applications of this scaling method can be found outside the realm of psychology, namely the chess rating system developed by the mathematician Arpad Elo (1978). His idea, to assign every active chess player with a number indicating playing strength, was adopted by the United States Chess Federation in 1960 and has been used by the Fédération Internationale Des Échecs (FIDE) since 1970 (Jonker, 1992). The chess rating system yields interesting implications for psychology, since the practical demands placed on such a system forced chess federations worldwide to overcome difficulties resembling the methodological restrictions that limit the use of paired-comparisons in science (Batchelder and Bershad, 1979). In this paper I seek to give an overview of the development of the Elo-rating system, including its statistical basis and suggestions for improvements.

2. Paired-Comparisons

2.1. Introduction

In paired-comparison scaling subjects typically have to evaluate pairs of stimulus-items and choose one over the other. Usually, these items are selected from a larger set, and the goal is to determine the overall position of each item on a dimension or *psychological continuum* that is meaningful for this set. In other words: the scale represents a characteristic C that applies to all items in the set. The meaning of the scale is determined by the question asked to the subjects. The scale can be used for two purposes: (a) the description of the underlying structure of the items; and (b) the construction of a measurement device based on the scale values.

Let i and j be two different items from a larger set or *universe* X consisting of N different stimuli. The outcome of the evaluation of i against j can be denoted as s_{ij}, with $s_{ij} = 1$ if $i > j$ (i is preferred over j) and $s_{ij} = 0$ if $i < j$. Mathematically, this can be described as a function

$$s\colon \{\, (i,j) \in X \times X : i \neq j \,\} \rightarrow \{0,1\}. \tag{1}$$

We can apply this to a chess context. In such a setting the subjects (chess players) in a sense *are* the stimulus-items, and the evaluations s_{ij} are the outcomes of games played between two players i and j. One of the largest problems for the Elo-rating system is constituted by the possibility of draws (ties).

The outcomes s_{ij} can be combined into scores $v \in \mathbb{R}$ on an interval scale that represents the characteristic C for all elements of X. Calculating v for all elements in X allows those elements to be sorted on C and ideally this induces an *irreflexive ordering* on X with relation $>$. An irreflexive ordering on a set X requires (a) *transitivity*—let a, b, and $c \in X$. If $a > b$ and $b > c$ then $a > c$; (b) *irreflexivity*—let $a \in X$ then $\neg(a > a)$. The second assumption implies *antisymmetry*—let a and $b \in X$. If $\neg(a > b)$ and $\neg(b > a)$ then $a = b$ (Doets, 1992).

An irreflexive ordering on a set of items represents the perfect outcome of a paired-comparison scaling experiment. Unfortunately neither criterium can be met

in most cases. Patterns such as $a > b$, $b > c$ and $c > a$ are found regularly, which violates the transitivity-principle. If in such a case only single outcomes are obtained for items a, b, and c it is impossible to order them according to $>$. However when multiple outcomes are obtained we can base our ranking on the *probability* p that $a > b$ when these items are evaluated against each other. Assuming that $i \neq j$ for any $i, j \in X$ the set X will *tend towards transitivity* when an infinite amount of results are obtained for each stimulus pair. Hence, X is a *probabilistic set*. Of course it is impossible to obtain an infinite amount of outcomes for all pairs, so the practical requirement is to obtain a large amount of outcomes.

Antisymmetry represents one of the main problem of the paired-comparison methodology—the problem of ties. According to (1) a choice between i and j is forced. However in a lot of situations ties are appropriate or even unavoidable (like in chess, where a large proportion of games end in a draw).

Finally, in most paired-comparison experiments the time- or space-order in which the stimulus-items are presented can influence the outcomes of the evaluations. These time- and space-order errors have to be attenuated when possible. In chess, errors of this kind occur in the form of the infamous *opening advantage for White*. I will elaborate on the transitivity principle here, the problem of ties and the opening advantage for White will be discussed below.

Let f_{ij} be the *frequency* of i being preferred over j and n_{ij} is the amount of outcomes obtained for the stimulus pair (i, j). Then

$$p_{ij} = f_{ij}/n_{ij} \tag{2}$$

is the *proportion* of times that i is preferred to j (or the probability that i will be preferred over j when being matched). Even though an irreflexive ordering will rarely occur for an entire set of paired-comparison data, the transitivity-principle holds for a subset $T \subset X$ with $T = \{a, b, c\}$ if from $p_{ab} = 1$ and $p_{bc} = 1$ follows that $p_{ac} = 1$ (Luce, 1959).

The probabilities p_{ij} can be transformed into scale-values v for each item in X such that v is larger when an item is preferred more often. This imposes a second requirement on paired-comparison scaling, namely that outcomes are obtained for all possible combinations of items from X. Otherwise v can not be calculated for all items and those v's that can be will not be accurate. Hence, in order to convert the outcomes obtained per pair to a scale value the following criteria have to be met: (a) results are obtained for all stimulus pairs; and (b) a large amount of results are obtained for each stimulus pair (Meerling, 1988).

Take two items i and j from X and determine p_{ij}. The scale value of i is denoted by v_i. If $p_{ij} = .5$ it is reasonable to say that $v_i = v_j$, since neither item is preferred over the other. Note that this is not the same as a tie: a tie is the outcome of an individual evaluation while the probability p_{ij} is based on many evaluations of which none may have resulted in a tie. In case $p_{ij} > .5$, then $v_i > v_j$ (or $v_i - v_j > 0$). Similarly if $p_{ij} < .5$, then $v_i - v_j < 0$. The relationship between p_{ij} and the *discriminal difference* $(v_i - v_j)$ can be expressed by a monotonically increasing function

$$p_{ij} = F(v_i - v_j) \tag{3}$$

provided that $v_i - v_j \to -\infty$ for $p_{ij} = 0$, $v_i = v_j$ for $p_{ij} = .5$, and $v_i - v_j \to \infty$ for $p_{ij} = 1$. Besides, $p_{ij} = 1 - p_{ji}$ which implies that F should be symmetric around $p_{ij} = .5$. Hence

$$F(v_i - v_j) = 1 - F(-(v_i - v_j)) = 1 - F(v_j - v_i) \tag{4}$$

Thurstone (1927/1994), and Bradley and Terry (1952) defined paired-comparison models based on different functions $F(v_i - v_j)$ that meet these criteria.

2.2. The Thurstone Case V Model

A classic paired-comparison scaling model is the Case V model of Thurstone's *Law of Comparative Judgment* (Thurstone, 1927/1994). This model is based on the *normal probability function*

$$\Phi(v_i - v_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(v_i - v_j)} e^{-\frac{1}{2}t^2}, \tag{5}$$

with $\sigma = 1$, which represents the cumulative *standard normal distribution of measurements* (see Figure 1).

Figure 1.: The normal probability function, $C = \sigma$ (reprinted from Elo, 1978)

This choice is a rather arbitrary, but sensible one. The normal probability function meets all the criteria for F mentioned above. Besides, outcomes are not always the same for a certain pair (i, j). If we assume that those fluctuations are random rather than caused by changing properties of i and j, the fluctuations will have a normal distribution, just like $v_i - v_j$.

At the same time it is very convenient to use the normal probability function, because it allows the use of z-scores to quantify $v_i - v_j$:

$$z_{ij} = v_i - v_j. \tag{6}$$

The calculation of v for all items i, j from X proceeds as follows: (a) all p_{ij} are determined using (2); (b) all p_{ij} are converted to z_{ij}-scores; (c) addition of all z_{ij} scores for each item i; and (d) division of this sum by k, hence:

$$v_i = \frac{\sum_{j=1}^{k} z_{ij}}{k}. \tag{7}$$

In other words: v_i is the average of all z_{ij} with $j = 1, ..., k$.

Note that the sum $\sum_{j=1}^{k} z_{ij}$ can be written as:

$$\sum_{j=1}^{k} z_{ij} = \sum_{j=1}^{k} (v_i - v_j) = kv_i - \sum_{j=1}^{k} v_j \tag{8}$$

where k is the amount of evaluations carried out. By substitution of (8) in (7) we can disclose more information about the scale values v_i:

$$v_i = \frac{kv_i - \sum_{j=1}^{k} v_j}{k} = v_i - \frac{\sum_{j=1}^{k} v_j}{k} \tag{9}$$

and $\sum_j v_j = 0$. Hence, the scale values v are expressed in their deviation from the mean (and $\mu = 0$). This is acceptable since the model is based on the *differences between scale values* which allows linear transformations.

(6) is a special case of the statement that

$$z_{ij} \sim v_i - v_j. \tag{10}$$

If (10) is our only requirement, the model does not only allow linear transformations, but multiplication with a constant c as well:

$$v_i{}' - v_j{}' = c(v_i - v_j), \tag{11}$$

and now differences in scale values can be transformed into probabilities according to a normal distribution with $c = \sigma$, or

$$p_{ij} = \Phi(\frac{v_i{}' - v_j{}'}{\sigma}). \tag{12}$$

If a probability $p_{ij} = 1$ is found for a certain $i,j \in X$, z_{ij} is no longer defined ($z_{ij} \to \infty$). This problem can be solved by substituting $p_{ij} = (n - \frac{1}{2})/n$, with n the amount of times that the stimulus pair (i, j) was evaluated. The underlying idea is that if n was chosen large enough, p_{ij} would never be equal to 1, hence we assume that $(n-1)/n < p_{ij} < 1$ and $(n - \frac{1}{2})/n$ is right in the middle. By analogy, $p_{ij} = \frac{1}{2n}$ is substituted for $p_{ij} = 0$. For experiments with $\#s_{ij} > 2$ consider Luce (1959).

2.3. The Bradley-Terry Model

A second model which is important for this discussion is the Bradley-Terry Model (Bradley & Terry, 1952). In this model Φ is specified as

$$\Phi(v_i - v_j) = \frac{1}{1 + e^{-(v_i - v_j)}}. \tag{13}$$

This is the cumulative *logistic* or *Verhulst function*.

Note that (13), just like (5) meets all the requirements for F mentioned above. The inverse

$$\Phi^{-1}(p_{ij}) = v_i - v_j = {}^e log \frac{p_{ij}}{1 - p_{ij}} \tag{14}$$

allows calculation of $v_i - v_j$ when p_{ij} is known. Calculation of individual values v_i proceeds similarly to the Thurstone Model Case V. Values v_i calculated according to the Bradley-Terry Model will be very similar to those calculated according to the Thurstone Model, since the normal probability function and the cumulative logistic function are almost identical, only the latter curve has higher tails (see Fifure 2).

Figure 2.: Standard normal and logistic distribution, $C = \sigma$ (reprinted from Elo, 1978)

According to Elo (1978) an important difference between these two functions is that scale values v_i calculated according to the Bradley-Terry model are defined on a ratioscale instead of an intervalscale. This is subject to a lot of discussion but since it is outside the scope of this paper I will not address this issue here. Nowadays the cumulative logistic function is usually prefered over the normal probability function because its derivative and inverse are easier to use.

(13) can be rewritten as

$$\Phi(v_i - v_j) = \frac{e^{(v_i - v_j)}}{1 + e^{(v_i - v_j)}}. \tag{15}$$

Now it becomes obvious that the logistic function used by Bradley and Terry constitutes an instance of the Rasch-Model in Item-Response theory (Meerling, 1988):

$$p_g(\theta) = \frac{e^{(\theta - b_g)}}{1 + e^{(\theta - b_g)}}. \tag{16}$$

Here $p_g(\theta)$ denotes the probability that an item g will be responded to in a specific way if the subject has value θ on a certain characteristic C. Similarly, b_g represents the value of this item g on C. In order to adapt this model to paired-comparison methodology, all that needs to be done is (a) replace the subject that answers questions with a certain item i and let $g = j$, then $\theta = v_i$ and $b_g = j$; and (b) define $p_g(\theta)$ as the probability that i will prevail over j, then $p_g(\theta) = p_{ij}$.

Again, $\Phi^{-1} = v_i - v_j$ can be expressed more leniently as

$$\Phi^{-1} \sim v_i - v_j \tag{17}$$

allowing both linear transformations and multiplication with a constant c, such that

$$p_{ij} = \Phi(\frac{v_i - v_j}{c}) = \frac{e^{(\frac{v_i - v_j}{c})}}{1 + e^{(\frac{v_i - v_j}{c})}}. \tag{18}$$

Besides, the base e in (14) can be replaced by any other number $a \in \mathbb{R}$, which results in the scale values being expressed in a different unit.

Luce (1959) interpreted the Bradley-Terry Model from the viewpoint of *Choice Theory*, which predicts the kind of choices someone would make in a certain situation. He assumed that ratio of the probabilities p_{ij} and p_{ji} remain the same, regardless of which and how many alternatives are added. This is the *independence-from-irrelevant-alternatives* principle. So

$$\frac{p_{ij}}{p_{ji}} = \frac{p_{ij}}{1 - p_{ij}} = \frac{p_i}{p_j}, \tag{19}$$

where p_i and p_j are the probabilities of i and j being picked out of the solution space s respectively, and this can be rewritten as:

$$p_{ij}(p_i + p_j) = p_i, \tag{20}$$

hence

$$p_{ij} = \frac{p_i}{p_i + p_j}. \tag{21}$$

Substituting (21) in (14) results in

$$v_i - v_j = {}^e log \frac{\frac{p_i}{p_i+p_j}}{\frac{p_j}{p_i+p_j}} = {}^e log(\frac{p_i}{p_j}) = {}^e log\, p_i - {}^e log\, p_j \tag{22}$$

Luce's extension of the Bradley-Terry Model creates a possibility to incorporate ties as a possible outcome of paired-comparison trials, which will be explained below.

3. Elo-Ratings

3.1. Introduction

The Thurstone model (Thurstone, 1927/1994) as well as the Bradley-Terry Model (Bradley & Terry, 1952) offers excellent possibilities for the development of skill-rating systems. In an environment where a group of individuals compete with eachother on a regular basis, rankings can be made by assigning a score or rating (v_i) to each participant according to either model. One of the largest applications of this kind can be found in the chess world. In the areas of Go and tabletennis similar systems are being introduced on a smaller scale.

Application of paired-comparison scaling to competitive environments requires specific difficulties to be addressed, some of the main problems being that (a) human ability changes over time due to learning processes and aging; (b) performances fluctuate due to "form of the day"; and (c) there is a massive turnover of participants in the universe X (in this case the chess world). Therefore ratings need to be updated from time to time in order to reflect the current skill level. The question is how often new ratings need to be calculated and to what extent older outcomes should be included in this new rating. If a new rating is based on new outcomes only, the rating change will, depending on the number of games played, be caused mainly by random fluctuations, instead of learning or aging processes, while inclusion of all previous outcomes will make real changes in ability almost invisible.

At the end of the fifties, Elo (1978) developed a chess rating system based on the Thursone Case V Model which has been adopted by chess federations worldwide. His rating system was not the first one to be launched; the first ratinglist was published in Germany by Hösslinger (1948) according to the Ingo-system. Other

examples are Harkness (1951) and Clarke (1958). However the Elo-system was much more successful, due to the fact that rating-differences $(v_i - v_j)$ and mutual winning chances are much more clearly related in this system than any other. Moreover, Elo was the first one to use computers for his calculations, which enabled him to rate a huge amount of players.

Unlike some other interval scales, such as temperature, the Elo-scale possesses reproducible fixed points nor depositable standard units. It is an open-ended floating scale. The *major interval* (unit) used is derived from statistics and probability theory. Individual performances are distributed normally. The standard deviation σ of individual performances is defined as 200 Elo points.

The initial transformation of the Thurstone Case V Model into the Elo-rating system consists of (a) a multiplication of all z_{ij} with $\sigma\sqrt{2}$; and (b) a translation over $M = 2000$. This allows chess ratings to be published in natural numbers (whole numbers > 0) without losing a great deal of accuracy.

Elo defines a *class interval* or *category interval* as the rating difference $(v_i - v_j)$ between the top and the bottom of a class. Individual performances are usually fluctuating around the average proficiency level v_i. A *class* consists of those players whose individual *ranges of performance* overlap at the time of measurement.

3.2. Formulas

If sufficient results s_{ij} are obtained for a group of players, *performance ratings* R_p can be calculated with the *performance rating formula*

$$R_p = R_c + D(p) \tag{23}$$

with R_c the average rating of the opposition and $D(p)$ the discriminal difference based on the percentage score p obtained against the opposition. The performance rating formula can be used to assign a rating to a new player based on his results against other players. For rated players, (23) indicates their level of play during one particular tournament.

The percentage score p is defined as

$$p = \frac{\sum_{j=1}^{k} p_{ij}}{k} \tag{24}$$

with p_{ij} the percentage score of player i against player j, and k the amount of different opponents. In tournament practice this is just the average amount of points obtained per game, thus (24) can be written as:

$$p = \frac{\sum_{j=1}^{n} s_{ij}}{n} \tag{25}$$

with s_{ij} the outcomes obtained in n games. This brings us to one of the main theoretical problems for the Elo-rating system and paired-comparison scaling in general: the solution space in chess is not limited to wins $(s_{ij} = 1)$ and losses $(s_{ij} = 0)$ but it includes draws (ties) as well. In chess, a draw has always been interpreted as halfway between a win and a loss, hence $s_{ij} = \frac{1}{2}$. Thus:

$$s\colon \{\, (i,j) \in X \times X : i \neq j \,\} \rightarrow \{0, \frac{1}{2}, 1\}. \tag{26}$$

See below for a discussion about the correctness of this assumption.

Upon introduction of the Elo-rating system, R_c was randomly set to 2000 and c was defined as $\sigma\sqrt{2}$, and $\sigma = 200$. Note that according to (23)

$$D(p) = R_p - R_c \tag{27}$$

and

$$R_p - R_c \sim v_i - v_j. \tag{28}$$

Hence it follows from (11) that

$$R_p - R_c = cz_{ij} \tag{29}$$

with $c = 200\sqrt{2} = 282.84$. By analogy, the *percentage expectancy curve* P can be obtained by multiplication of the normal probability function (5) with $\frac{1}{200\sqrt{2}}$:

$$P(z_{ij}) = \frac{1}{200\sqrt{2}}P(R_p - R_c) = \frac{1}{800\sqrt{\pi}}\int_{-\infty}^{R_p - R_c} e^{-\frac{1}{2}t^2}. \tag{30}$$

If a rating is to be calculated for a new player, R_c can have any value > 0 depending on the ratings of the opponents. For a new rating to be reliable it needs to be based on at least 30 games against rated players. Players with more than 9 but less than 30 rated games receive a *provisional rating* which is their performance rating R_p calculated over the games played so far.

In *periodic* rating systems, such as the BCF-rating system (Clarke, 1958), (23) is used to calculate new ratings for all players on a regular basis, using previous ratings of the opponents to determine R_c.

However, most rating systems are conducted on a *continuous basis*, allowing previous results (represented by the old rating R_o) to excert some influence on the current rating R_n. Assume that for a certain player i the old rating R_o is based on an amount of games N_o with $N_o \leq 30$ and this player participates in a new event resulting in a performance rating R_p based on a new sample of N games and $R_p \neq R_o$. In order to combine R_o and R_p into a new rating R_n we should attenuate previous results in favor of the newer ones while making sure that random fluctuations do not give rise to a false rating statement.

The desired result can be obtained by weighting R_o and R_p by the amount of games they are based on (N_o and N respectively) while *pretending* that the total amount of games R_n is based on equals N_o. Hence

$$R_n = \frac{R_o(N_o - N) + R_pN}{N_o} = R_o + (R_p - R_o)\frac{N}{N_o} \tag{31}$$

and

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

If we assume that the average competition R_c in the sample N_o does not differ from that of N, ΔR can be written as

$$\Delta R = [D(p) - D(p_o)]\frac{N}{N_o} \tag{33}$$

and p_o represents the percentage scored in N_o. In case p and p_o do not differ to a great extent ($|p - p_o| \leq 3\sigma$) they can be expressed in terms of the percentage expectancy curve with slope S. Substitution in (33) gives

$$\Delta R = [\frac{1}{S}(p - p_o)]\frac{N}{N_o}. \tag{34}$$

For most of the percentage expectancy curve, $S = \frac{1}{4\sigma}$, hence

$$\Delta R = \frac{4\sigma(Np - Np_o)}{N_o}. \tag{35}$$

This last equation can be rewritten as a function of the *expected outcome* W_e. When two rated players i and j are competing, the expected outcome of the game can be defined as $W_e = p_{ij}$. The real outcome of the game W however may differ from W_e. If W=W_e the rating difference $v_i - v_j$ (with v_i the rating of player i) correctly reflects the difference in ability between the two players, but if $W \neq W_e$ the ratings v_i and v_j need to be corrected until $W = W_e$. This can be extended to the case of a player competing in a tournament. His score in the tournament W (defined as the number of wins and half the number of draws) may not be the same as the expected score W_e, which indicates that the rating of the player needs to be adjusted.

$$W = Np \tag{36}$$

and

$$W_e = \sum_{j=1}^{k} p_{ij} = Np_o \tag{37}$$

when k stands for the number of rounds in the tournament. Substitution of (36) and (37) in (35) gives

$$\Delta R = \frac{4\sigma}{N_0}(W - W_e). \tag{38}$$

Now

$$R_n = R_o + \Delta R, \tag{39}$$

and from substitution of (38) in (39) follows that

$$R_n = R_o + K(W - W_e) \tag{40}$$

with $K = \frac{4\sigma}{N_o}$.

Hence the impact of the difference $W - W_e$ depends on the *rating point value* K, which is the maximum amount of rating points that can be won in one game. A high K-factor gives more weight to new results while a low value increases the influence of earlier performances.

Note that (40) enables calculation of a new rating R_n after a single game as well as after a tournament. In the first case W and W_e represent the actual outcome of the game and the probability p_ij respectively while in the second case W and W_e denote the total score in the tournament and the expected score as follows from (37).

Most federations are currently changing their rating calculation methods from the second method (calculation after a tournament or even an entire rating period consisting of several tournaments) to the first. One of the advantages of the first method is nicely illustrated by the following example. Imagine that the author of this paper (KNSB-rating 2027 of September 1998) plays 2 games during a rating period, one of which against a 2800 player and the other one against someone rated 1100. If things proceed normally I will lose against the strong player and beat the weak one. If my rating is updated after each game I will end up with a new rating of $R_n = R_o = 2027$ since my expected scores are 0 and 1 respectively and that is exactly how I performed. However, if my new rating is calculated after the 2 games my new rating will be

$$R_n = 2027 + K(1 - 1.22) = 2027 - 0.22K$$

and unless my K-factor is 0 I will lose rating points. Since this is rather unfair, it is recommended to update ratings after each game played, especially if rating differences are very large (as is the case in the example).

3.3. Chess Ratings according to the Bradley-Terry Model

The same definitions can be made according to the Bradley-Terry Model (Bradley & Terry, 1952). The expected score of player i against player j if the rating difference $(v_i - v_j)$ is known, is defined as

$$p_{ij} = \Phi(\frac{v_i - v_j}{400}) = \frac{(10^{\frac{v_i - v_j}{400}})^{s_{ij}}}{1 + 10^{\frac{v_i - v_j}{400}}}. \tag{41}$$

In this version of the Bradley-Terry Model scale-values are expressed on a decimal scale with $\sigma = 400$. This is the basic formula for the rating system of the United States Chess Federation.

3.4. Ties

One of the largest problems in paired-comparison methodology is constituted by ties. In most cases ties can be avoided simply by not allowing them, but in chess the possibility of draw is inherent to the rules of the game. Hence, not allowing draws would change the nature of the game completely. The following game-situations lead to a draw: (a) stalemate—the player to move cannot dispose of any legal moves and is not in check (otherwise it would be mate); (b) insufficient material—the material left on the board is insufficient for one of the players to be mated (for instance when the Kings are the only pieces left on the board); and (c) the *fifty moves rule*, which applies to situations when 50 consecutive moves have been played without any pawns being moved or any pieces captured. Normally games are also drawn if both contestants *agree* on a draw, but theoretically this option can easily be excluded.

If any of those situations occur and draw is not allowed, the game would be *permanently undecided*. This is exactly how ties were dealt with in the old days when tournaments did not exist and the legends combatted eachother in matches (a series of more than one game against the same opponent with alternating colors). The first player to win a certain amount of games would win the match, and draws were not counted in the score. (Only recently has this practice for match play been abandoned by FIDE, for the simple reason that the large amount of games resulting in a draw made championship matches incompatible with the contemporary demand for quick decisions.) However, this method is not suited for tournament play and therefore a draw was defined as halfway between a win and a loss, with $s_{ij} = s_{ji} = \frac{1}{2}$. This method proved perfectly satisfactory in practice, but this does not eliminate the theoretical problems connected with ties.

Assigning both players with the same score implies that both players put up equally strong opposition. It is questionable though if this is in agreement with reality. Glickman (1998) proposes a solution for this dilemma by considering the probability ${p_{\frac{1}{2}}}^2$ of two consecutive draws against the same player as equivalent to a win followed by a loss (or a loss followed by a win). Hence

$${p_{\frac{1}{2}}}^2 = p_{ij}(1 - p_{ij}) \tag{42}$$

and

$$p_{\frac{1}{2}} = \sqrt{p_{ij}(1 - p_{ij})}. \tag{43}$$

In this model p_{ij} is the *expected score* of player i in his game against player j.

According to the *independence from irrelevant alternatives* principle (Luce, 1959) the ratio of the expected scores p_{ij} and p_{ji} remain the same when a new alternative ($p_{\frac{1}{2}}$) is added to the model. Hence

$$p_{ij} = 1 \times p_i + 0 \times p_j + \frac{1}{2} \times p_{\frac{1}{2}} = p_i + \frac{1}{2} p_{\frac{1}{2}} \tag{44}$$

and p_i, p_j are the probabilities that respectively player i or j will actually win the game. If p_{ij} is known, $p_{\frac{1}{2}}$ follows from (43) and $p_j = 1 - (p_i - p_{\frac{1}{2}})$.

This statement is based on the assumption that the probabilities p_{ij} for 2 consecutive games are independent. The problem with this assumption is that independence of probabilities implies that the properties of the players such as playing strenght and fighting spirit are static. So first of all it implies that players do not learn from their previous games. But if playing strenght can not change between games then when *will* it change? Second of all, the result of a game may affect the player's attitude for the next game. Even though it is unclear *how* fighting spirit excerts its influence on future play, it is unlikely that it is not of any importance at all. However the influence of above factors will be marginal, and if the object is to avoid addressing the problem of ties, Glickman's proposal deserves serious consideration.

However there is another problem concerning Glickman's idea that will be more difficult to discard, being that overall, draw occurs more frequently among Grand Masters than lower level players. When both players have the same rating ($v_i - v_j = 0$), $p_{\frac{1}{2}} = 0.5$. This outcome certainly seems to be right or even on the low side at Grand Master level (considering the large amount of games being undecided there) but it does not seem to apply to lower levels, where a substantial amount of games is decided by *blunders* and draw occurs less frequently. See Jonker (1992) for a discussion of 3 models for situations with a high draw ferquency, a low draw frequency or no draws at all.

Maybe it is possible to correct for this "luck" factor by multiplying (43) with a function $b(r)$ which reflects the influence of sheer chance on game outcomes based on a parameter r:

$$p_{\frac{1}{2}} = b(r)\sqrt{p_{ij}(1 - p_{ij})} \tag{45}$$

with $0 \leq b(r) \leq 2$ and increasing $b(r)$ s reflecting a decreasing influence of random fluctuations. Now $p_{\frac{1}{2}} = 1$ if both contestants play chess at perfection-level (assuming that the opening advantage for White, to be discussed in the next subsection, is only a practical and not a theoretical one).

Note that this parameter could contain information about the rating-level as well as personal characteristics of the players which could possibly increase or decrease the likelihood of a draw, such as playing style, routine, fighting spirit, age and gender. Unfortunately it will be very difficult to determine the nature of the influence of these characteristics, let alone quantify them, because they typically covary intra- and interpersonally and all of them (except gender) are not stable over time.

Last but not least it is questionable if two consecutive draws provide the same amount of information about the rating difference $(v_i - v_j)$ as a win followed by a loss. In short it can be concluded that Glickman's idea is interesting, but only applicable to Grand Master chess, and even then with caution.

3.5. The Opening Advantage for White

In the contemporary chess world it is not uncommon to hear people claim that one should win with White and draw with Black. The bias for White is caused

by the fact that the one who moves first has the *initiative* (which is an advantage) and constitutes an error of the same type as the *time-order* and *space-order* errors frequently encountered in paired-comparison scaling.

Jonker (1992) suggests to add 25 points to White's rating and subtract 25 points from Black's in order to correct for color. This implies that color adds 50 points to the rating-difference $v_i - v_j$ in White's favor. I think however that this is very impractical since as a consequence all players would have to use different ratings for White and for Black. Besides, White's opening advantage is less prominent on lower levels, being non-existent among beginners. Therefor the compensation for color should be different for players of different skill levels.

A different way of accounting for the opening advantage is to alter the way points are awarded for a chess game, making a Black win or draw worth more rating points than a White win or draw. For 2 players i and j of equal strength, p_{ij} (with i is White and j is Black) would be 0.57 (reflecting Jonker's 50 points, see above) instead of 0.5. In order to compensate for this advantage, all s_{ij} should be translated over an interval $d = 0.57 - 0.5 = 0.07$ in Black's favor. This correction puts p_{ij} back at 0.5:

$$p_{ij} = 0.57 \times 0.43 - 0.43 \times 0.07 = 0.5$$

Note that a White loss is rewarded a negative score of -0.07 in this model (and a Black win a positive score of 1.07)! Of course this does not solve the problem that the size of the opening-advantage is dependent on level of play and besides, I think that altering the way games have been scored for ages is highly undesirable. In this case the theoretical gains certainly do not outweigh the practical losses.

3.6. Inflation and Deflation

3.6.1. Introduction

I would like to remind the reader that scale-values v_i are not intrinsically meaningful. Only the differences $(v_i - v_j)$ between scale-values carry information. Elo-ratings are measured on an open-ended floating scale and they surely behave as such. It is not difficult to understand that since the introduction of the Elo-rating system, the range of scale-values has been expanding constantly. This is mainly caused by tournament chess becoming increasingly popular as well as the use of Elo-ratings becoming more widespread. In the past Elo-ratings were only calculated for the happy few who were strong enough to participate in world class tournaments. Nowadays ratings are available for everybody. As the amount of players in the pool increases, the probability of some of them being extremely weak or extremely strong increases as well. The expanding range of Elo-ratings is a normal phenomenon and does not pose any problems for the system.

More complicated are the effects of (a) players entering and leaving the rating pool (*turnover*); and (b) the influence of *subpools* (such as scholastic chess players in the United States) on chess ratings because these factors, by causing inflation as well as deflation, put the *integrity of the rating system* (Elo, 1978) under pressure. The integrity of the rating system indicates to which extent a given rating v_i reflects the same level of chess proficiency from one point in time to another and across subpools. It is important for a rating system to remain more or less stable over longer periods of time in order to be able to handle people who play with different frequencies or in different subpools. It is difficult however to establish the net effect of the various inflating and deflating factors.

In order for a rating system to remain stable, the same amount of rating points should be available per person in the rating pool at different points in time. In other words: the average rating should remain the same. The only way to guarantee the maintenance of the rating average is to constantly rescale all ratings whenever one player enters or leaves the rating system. however, this way it would be possible to gain or lose rating-points without playing games, or to lose rating points despite the fact that the rated game was won. Hence most rating systems do not operate that way.

3.6.2. Turnover

Players typically leave the rating pool with a higher rating than they entered it, simply because they improved their play during their chess career. It is true that chess proficiency will decline as well for people who stay in the rating pool long enough, but this does not compensate for the rating deflation caused by people who do not.

If R_i is the initial rating assigned to a player upon entering the pool and R_q the rating upon retirement, an amount of $R_i - R_q$ rating points is taken away from the rating pool. Since R_q is not known in advance, all attempts to anticipate this loss by adding points to the system should be statistical in nature (Elo, 1978). How it is possible to ensure that those extra points are given to the players who are most entitled to them is not clear to me.

In the FIDE-rating system part of this deflation is counteracted by the inflation caused by players who leave the rating pool because their rating drops below the 2000-level. Most of these people entered the pool with a rating which was too high, and these points stay around when the people who brought them in are already gone.

Deflation is not equally prominent at different levels of play. Players whose chess playing capacities are limited tend to leave the rating system shortly after entering causing greater turnover without having much impact on the rating system. Stronger players typically stay in the rating pool for longer periods of time, and it is these stronger players who are causing the deflation.

3.6.3. Subpools

Inflation and Deflation do not only occur in the rating pool as a whole but also within *subpools*. A subpool is a fraction of players included in a rating system who over longer periods of time only compete with each other and not with people from outside the group. In mathematical terms, one could say that a subpool is a subset $T \subset X : T \cap X \backslash T = \emptyset$ (with X universe, or the rating system), provided that all elements $t \in T$ can be $\in X \backslash T$ as well, when they compete with elements $x \notin T$.

A striking example of such a subpool are junior players (under age twenty), who usually compete in their own tournaments. In a lot of chess federations junior tournaments are not rated, because of the fact that most of these tournaments use a different time control than "regular" events and the relatively unreliable performance of young players, whose proficiency can improve dramatically from one tournament to another. The USCF however rates all junior (scholastic) events and this accounts for a dramatic demonstration for what can happen when deflation is not corrected for.

The majority of the USCF scholastic chess players are rated < 1000, whereas in other rating systems, such as the Dutch KNSB rating system, ratings of that kind are rare because novices are not usually rated so early in their chess career. Chess players in the USCF receive a provisional rating after 4 games (9 games in the KNSB). The problem lies in the fact that the average rating in the scholastic subpool is several hundreds of points lower than the average rating for the entire

system. Most of the scholastic players improve dramatically during their stay in the scholastic subpool, but this is not always reflected by their ratings since the amount of rating points to be divided is rather small. In other words: by competing with each other only, they keep their own ratings low. Hence a lot of scholastic players are underrated. Within the scholastic subpool, ratings may still have a reasonable predictive value (although not as good as regular ratings, since most scholastic players are beginners as well), but as soon as scholastic players enter in regular tournament play the difference $v_i - v_j$ between a scholastic and a non-scholastic player becomes meaningless. The scholastic player will start winning a lot of rating points quickly, for he (or she) is underrated and this causes massive deflation of the rating system, a deflation that is more pronounced in the USCF system than in most other systems.

At the other end of the rating scale, at Grand Master level, a similar subpool effect causes some inflation. Grand Masters (especially so called Super Grand Masters with FIDE-ratings > 2650) tend to participate in small, closed tournaments in which only Grand Masters (and sometimes also International Masters) participate. This protects them from losing rating points against promising players who are underrated (something that happens to other players on a regular basis) and henceforth artificially keeps their rating high. This effect is much smaller though than the deflating effect caused by beginners, but it does take away some strength from Kasparov's claim that he is the strongest player ever on the face of earth (one of his main arguments is his current FIDE-rating of 2815).

Altogether the subpool-phenomenon shows that not only it is important to obtain a large amount of outcomes for each player but also to have a reasonable diversity among the opposition.

3.6.4. Solutions

In order to keep deflation under control a chess federation can take the following precautions: (a) when ratings are updated, (provisional) ratings should be calculated first for new players, then existing provisional ratings should be evaluated based on their results against all other players including the new players (whose new ratings are used in the calculations), and finally all established ratings can be updated. This isolates new players as a potential source of deflation; (b) if subpools do exist within a rating system (such as the USCF scholastic players), it is advisable to rescale their ratings before the established ratings are updated; (c) in order to protect established ratings against random fluctuations, the K-factor can be varied as a function of the amount of games played or the rating of a player (more games played or a higher rating needs a lower K-factor); (d) players who show extraordinary performance compared to their rating can be awarded bonus rating points or a lower K-factor. These so-called outlayers can be detected with a *z-M test* (Elo, 1978); and (e) new ratings from a first cycle of calculations to update old ratings can be used as input for a second cycle (and then again for a third etc.) in order to stabilize the new ratings somewhat. This *iterative* method results in an optimal distribution of available rating points among the players.

The composition of a rating pool as well as the ways in which ratings are calculated can differ greatly from federation to federation. Therefore it is not advisable to use ratings from one rating system in order to update ratings in another (which is tempting to do for international tournaments). Means and standard deviations are usually not the same so that the differences $(v_i - v_j)$ are meaningless if v_i and v_j are not from the same system. (Likewise it is not possible to calculate sums with numbers from different numerical systems without performing some transformations first.) Mindlessly mixing ratings from different systems can result in serious contamination. A lot of chess federations however incorporate results obtained for

FIDE-rating in their own ratings as well, in order to keep their own rating system and the FIDE-ratings more or less comparable. This way, national ratings can be calibrated to FIDE-ratings.

4. Discussion

Despite all the methodological imperfections of the Elo-rating system it has stood up well in chess tournament practice for almost 50 years now. It turns out that restrictions such as (a) ties; (b) the problem of sparse data; (c) new players entering the system; and (d) dynamic changes such as learning processes (see also Batchelder & Bershad, 1979, for a detailed mathematical discussion of these and other problems) can be overcome with reasonable success. This could have interesting implications for paired-comparison methodology, especially concerning the forced-choice paradigms. Allowing ties does not seem to pose insurmountable problems on the chess rating system which indicates that it should not do so either in other instances of paired-comparison scaling. However it is important to stay aware of the complications that can arise when requirements necessary to keep a rating system healthy and reliable are violated too lightheartedly, as is shown by the scholastic chess example given above.

Current research topics in this area focus on the influence of covariate information such as age, playing routine, style and proficiency level on the reliability of the rating system. Glickman (1998) proposes to incorporate these *uncertainties* in the $K - factor$:

$$K = \frac{q}{\frac{1}{\delta^2} + \frac{1}{\sigma^2}} \tag{46}$$

with δ^2 the variance of the normal approximation to the marginal likelihood of the correctness of the ratings v_j of the opposition and σ^2 the variance of the proficiency level of the player (as reflected by the rating v_i). Hence the idea is that the impact of the outcome of a game should be small if the variance of the rating of the opponent as well as own playing strength is large. For a detailed discussion of this proposal see Glickman (1998).

Another interesting example of the use of Elo-ratings in psychological research is a current research project by Van Der Maas, Wagenmakers, De Blécourt and Kamminga (in preparation) where Elo-ratings are used to indicate the difficulty of chess positions.

Whereas the Elo-rating system was derived from the paired-comparison scaling methodology as it is used in psychology, it nowadays sets the example on how to deal with methodological problems that have hampered the use of paired-comparison techniques in science. Not only it has evolved into one of the best ranking systems in the world today, it also is a striking example of how science and real-life applications can (and should) interact.

5. References

Batchelder, W. H., & Bershad, N. J. (1979). The Statistical Analysis of a Thurstonian Model for Rating Chess Players. *Journal of Mathematical Psychology, 19,* 39-60.

Bradley, R. A., & Terry, M. E. (1952). The Rank Analysis of Incomplete Block Designs. 1. The Method of Paired Comparisons. *Biometrika, 39,* 324-345.

Clarke, R. W. B. (1958). British Chess Federation Grading System. *BCF Yearbook,* London: BCF. (From: Elo, A. E. (1978). *The rating of chess players past and present,* New York: Arco Publishing.)

Doets, H. C. (1992). *Logica en Verzamelingen*, Amsterdam: F.W.I., Universiteit van Amsterdam.

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

Glickman, M. E. (1998, unpublished manuscript). Becoming a chess master—the development of a rating system for tournament chess players. .

Harkness, K.(1956). *Official Blue Book and Encyclopaedia of Chess*, New York: McKay. (From: Elo, A. E.(1978). *The rating of chess players past and present*, New York: Arco Publishing.)

Hösslinger, A.(1948, April). Ingo System. *Bayerischen Schachnachrichten*. (From: Elo, A. E.(1978). *The rating of chess players past and present*, New York: Arco Publishing.)

Jonker, H.(1992). *Het Nederlandse Elo-rating boek: Elo-ratings en het KNSB-ratingsysteem*, Venlo: Uitgeverij Van Spijk B.V.

Luce, R.,D.(1959). *Individual Choice Behavior*, New York: Wiley.

Meerling(1988). *Methoden en technieken van psychologisch onderzoek. deel 2. Data-analyse en psychometrie*, Meppel: Boom.

Thurstone, L. L.(1994). A Law of Comparative Judgment. *Psychological Review,101*, 266-270. (Original work published 1927)

Van Der Maas, H. L. J., Wagenmakers, E. J., De Blécourt, S., & Kamminga, J. (in preparation). Analysis of Chess Ability: a Psychometric Approach.