ANIMAL BEHAVIOUR, 2001, **61**, 489–495
doi:10.1006/anbe.2000.1571, available online at http://www.idealibrary.com on IDEAL®

Elo-rating as a tool in the sequential estimation of dominance strengths

PAUL C. H. ALBERS & HAN DE VRIES
Ethology and Socio-ecology Group, Utrecht University

(Received 18 May 2000; initial acceptance 18 July 2000; final acceptance 11 September 2000; MS. number: SC-1189)

Many methods of dominance rank ordination were recently reviewed by de Vries (1998). Overall, two types of method for finding a dominance rank order can be distinguished. In one group of methods some numerical criterion, calculated for the dominance matrix as a whole, is minimized (or maximized) resulting in a reorganized matrix for which this criterion is smallest (or largest). The result produced by each of these methods is a rank order of the individuals, that is, the most plausible one relative to the specific criterion used, and given the dominance encounters observed. This group includes methods developed by Slater (1961), de Vries (1998), McMahan & Morris (1984), Brown (1975), Bossuyt (1990), Crow (1990) and Boyd & Silk (1983). The second class of methods aims to provide a suitable measure of individual overall success in the group, from which a rank order can be directly derived. Measures that have been put forward for this purpose include: number of individuals dominated; proportion of total encounters won; Clutton-Brock et al.'s (1979) index of fighting success; David's (1987, 1988) score; and Jameson et al.'s (1999) score. As yet, however, none of these success measures appears to be generally accepted (see also de Vries & Appleby 2000).

All these methods start by observing behaviour for a certain period of time after which the outcomes of the dominance encounters are arranged in a matrix. When sufficient interactions between the contestants have been observed, a rank ordination method is used that yields a dominance order that is presumed to have existed during the whole observation period. Basically this means that it is assumed that specific interactions between two individuals reflect the dominance order rather than influence it.

In this paper we present the Elo-rating method which provides sequential estimations of individual dominance strengths based on the actual sequence of dominance interactions. From the values of the individual Elo-ratings an estimated rank order can be derived at any moment in time. Elo-rating was developed and subsequently named after Arpad Elo (1961, 1978). It is intended and still used as a fair method for ranking chess players. The Elo-rating calculation procedure is based on the assumption that the chance of A winning from B is a function of the difference in current ratings of the two contestants. After each contest the Elo-ratings of the two contestants are updated in proportion to the deviation of the actual outcome (win, loss or tie) from the expected outcome for each of the two contestants. The expected outcome for a contestant is based on the rating difference of the two contestants at the moment of the contest. The winner's rating increases (and the loser's rating decreases) in proportion to the deviation from the expected outcome. As outstanding features of the Elo-rating method in comparison to other methods of estimating dominance strength, we mention that it is independent of the number of contestants (which may vary over time), it takes the sequence of interactions into account, and gives a continuous update so that the process of dominance strength acquisition can be followed from interaction to interaction. Our main aim is to present Elo-rating as a method for the sequential estimation of dominance strengths. However, from a different perspective it is also possible to consider the Elo-rating updating process as a model of the way in which dominance is generated within a group. The underlying model here is based on the positive (negative) reinforcement of some internal variable when an individual wins (loses) a dominance interaction. We also briefly discuss the application of Elo-rating in a simulation modelling context.

Correspondence: P. C. H. Albers, Dikbosstraat 56, 7814 XP, Weerdinge, The Netherlands (email: palbers@xs4all.nl). H. de Vries is at the Ethology and Socio-ecology Group, Utrecht University, P.O. Box 80.086, 3508 TB Utrecht, The Netherlands.

The Elo-rating Method

In this section we present the Elo-rating method in some detail (see e.g. Elo 1961, 1978; Batchelder & Bershad 1979).

The rating of the winner of the contest is increased by an amount that depends on the chance of winning: the amount is small if the chance of winning is high and vice versa. Thus, a win by a high-rating individual (A) over a low-rating individual (B) increases A's rating only by a small value and decreases B's rating by the same small value. For example, if $A_{(\text{Elo-rating } 1200)}$ meets $B_{(\text{Elo-rating } 1000)}$, the difference 200 (1200 − 1000) corresponds to a chance

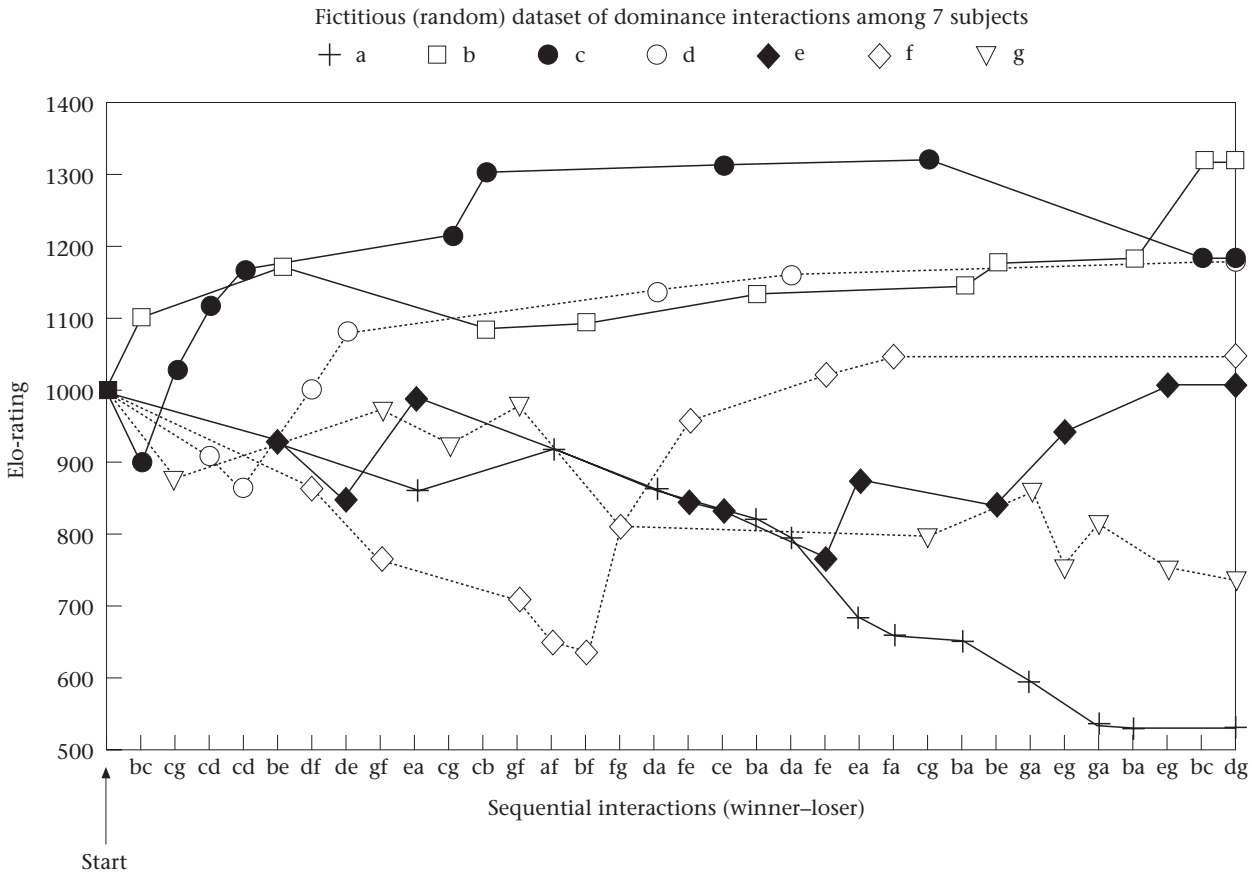

Figure 1. Sequential estimation of Elo-ratings (k=200) of seven subjects based on a fictive, random sequence of 33 dyadic interactions. Note that the horizontal axis is not a time axis but gives the sequence of events.

that A wins of 76% ($p_{\text{A wins from B}}$=0.76) (see Appendix). If A wins, its score is 1 instead of the expected 0.76 and the rating of A is subsequently updated according to the formula:

$$\text{New rating A}=\text{Old rating A}+(1-0.76)\times k$$

in which k is a constant. If we set k to 100 in these examples the rating updates to:

$$\text{New rating A}=1200+0.24\times 100=1224$$

For B, whose score is 0 since it lost, the same formula is applied:

$$\text{New rating B}=\text{Old rating B}+(0-0.24)\times k$$
$$=1000-0.24\times 100=976$$

If, on the other hand, B had won, against the odds, as its initial rating was lower than A's, its increase in rating would have been larger: its score from this single contest would have been 1 instead of the expected 0.24 and the ratings of B and A would be updated to:

$$\text{New rating B}=\text{Old rating B}+(1-0.24)\times k$$
$$=1000+0.76\times 100=1076$$

$$\text{New rating A}=\text{Old rating A}+(0-0.76)\times k$$
$$=1200-0.76\times 100=1124$$

Finally, if the contest between A and B had ended in a draw (tie), this would have been to the disadvantage of the higher ranking individual, as according to the same formula the expected outcome is now matched to the outcome of the tie (0.5) and the ratings would have been updated to:

$$\text{New rating B}=\text{Old rating B}+(0.5-0.24)\times k$$
$$=1000+0.26\times 100=1026$$

$$\text{New rating A}=\text{Old rating A}+(0.5-0.76)\times k$$
$$=1200-0.26\times 100=1174$$

Figure 1 illustrates how Elo-ratings change from an assumed all subjects are equal start (rating 1000). To get a feel for the Elo-rating updating method it is instructive to compare the early win of c from g (2nd interaction in the sequence) to the win of c from g later on (24th interaction). After the 2nd interaction the rating of c rather strongly increases (and that of g equally strongly decreases), whereas after the 24th interaction the win of c from g results in only a minor increase of c's rating and an equally small decrease of g's rating.

The constant k in this example is set arbitrarily to 200. Note that k influences only the amount by which the rating of a player increases or decreases after a single interaction, not the average rating of the group. A

different choice for the value of *k* results in exactly the same graph with only the scale of the vertical axis changed. In chess, the value of this constant *k* has been made dependent on the number of contests played, larger for a beginning player, smaller later on, thereby enabling a fast rising of a player to the appropriate rank but a slower re-positioning once its rating is a good estimate of its actual current strength. We already note here that, when one uses Elo-rating as a model for generating dominance in a group, the choice of *k* can have interesting implications; this will be discussed below.

As ordination is based on single interactions between individuals a ranking can be obtained independently of the specific period of observation. Indeed, a rank ordination can be made at any moment in time (an initial minimum number of interactions must have been observed, mainly depending on group size). This also means that the development or acquisition of rank in time can be examined from interaction to interaction. Since the Elo-rating allows for a linear ordering of all contestants, the use of Elo-rating implicitly assumes that a linear rank order exists at any moment in time. Of course the ratings obtained from the sequence of wins and losses up to some moment may differ (more or less) from the presumed actual strengths. In fact, the Elo-rating just continually keeps on estimating these presumed actual strengths of the individuals.

Ordination is also independent of the number of contestants; new contestants can be easily integrated and disappearing contestants do not make it necessary to break the data set up in different partitions. The rating itself is at the interval-scale level, so differences between the ratings are numerically meaningful.

Batchelder & Bershad (1979) assumed that Elo-rating would prove useful to track temporally changing ability parameters. In particular they showed mathematically that the strength of the opponent chosen does not influence the position in the ordination. In other words, individuals that avoid strong opponents but defeat weaker opponents do not surpass the stronger ones on the ordination scale if their actual strength of the ability in question does not justify that. This is particularly interesting as the lack of interactions between certain individuals in an agonistic dominance matrix can sometimes be a problem for deriving a rank order. Next we present two examples to show that the Elo-rating method is a useful and powerful tool to complement the existing methodology.

Examples

Example I

To show the fruitfulness of the Elo-rating method we applied it to a sequence of agonistic dominance interactions performed by 13 female rhesus macaques, *Macaca mulatta*, that were caged together for the first time in the course of a resocialization programme for previously solitarily housed monkeys. Behavioural observations were done by A. L. Louwerse and K. Berkhout (unpublished data). Initially all monkeys were assigned an Elo-rating of 1000. The constant *k*, which controls the rate of changing, was set to 20. With the Elo-ratings available at any point in time, the observations can now be analysed in more detail than was possible before. For instance, a close look at the rating differences in relation to the type of agonistic behaviours performed revealed that biting occurred only when the difference in rating exceeded 80. Since the difference in rating is supposed to reflect the chance of winning, this finding suggests that individuals engage in escalated fighting only if they consider their chances of winning sufficiently high (in this case about 60%). This may seem a somewhat trivial conclusion, but it could not be previously quantified so clearly. This example shows that Elo-rating enables the testing of refined hypotheses related to possible differences in fighting tendencies.

Another useful feature of the Elo-rating method is that the changes in dominance strength of each individual can be followed from interaction to interaction. This enables the precise investigation of how individuals acquire and/or maintain their positions in the dominance rank order.

Example II

We made a number of artificial data sets which each contained a large number of dominance interactions between a group of individuals in a distinct sequential order. We created an underlying linear hierarchy by making a list in which the chance of winning was established beforehand for each of the two contestants in all possible pairs. Subsequently for each successive interaction the winner was determined by a random draw with the chance of winning according to this previously made list.

After this sequence of dominance interactions and their outcomes were established, all individuals were given an initial Elo-rating of 1000 and the constant *k* was set to 20. From the given sequence of interactions and their outcomes, Elo-ratings were calculated for each individual after each interaction. For each of the data sets the Elo-ratings obtained at the end of the sequence correlated highly with the rank order found by the I&SI method, a method that derives a rank order from a set of dominance relationships, such that the number of inconsistencies, I, and, subsequently, also the total strength of these inconsistencies, SI, is minimized (de Vries 1998). Correlations between the two rank orders were higher the more the data set looked like what we considered a 'real' data set: in particular if we made low-ranking individuals initiate fewer interactions towards high-ranking individuals than the other way around. This example shows that if a stable dominance hierarchy is present and the outcomes of the interactions reflect this dominance order, the Elo-rating method and the I&SI method yield similar results, as expected.

Elo-rating as a Model for Generating Dominance in a Group

An interesting phenomenon in some animal species is that after an individual has won a contest its chances of

Table 1. Rank orders generated by a simulation model in which Elo-rating is used for generating dominance

	k=20		k=100		k=200	
	Order	Rating	Order	Rating	Order	Rating
	A	1740	A	2863	M	3768
	B	1369	E	2477	D	3180
	C	1262	K	2172	C	2562
	D	1162	H	2012	L	2260
	E	1106	B	1441	E	1838
	F	1077	D	1405	F	1492
	G	1074	C	811	B	998
	H	977	J	702	A	552
	I	814	I	630	G	300
	J	790	L	128	H	−196
	K	640	M	−306	I	−676
	L	571	F	−440	K	−1254
	M	418	G	−895	J	−1824
Linearity (Landau's h)		h=1.00 P<0.0001		h=0.99 P<0.0001		h=1.00 P<0.0001
Spearman rank correlation between Elo and I&SI order		r_s=0.85 P<0.001		r_s=0.87 P<0.001		r_s=0.91 P<0.001

winning a following contest will increase, even with another individual. This was investigated by Chase and coworkers (1982, 1985, 1987, 1994). If the Elo-rating updating method is taken as a model of the way in which dominance is generated in a group, it can be used for a relatively simple theoretical examination of this phenomenon. The essential difference between the use of Elo-rating as a basis for a simulation model and its use as a method of estimating the dominance strengths in a real observed data set is that in the model the outcomes of dominance interactions are generated according to the current difference in ratings of the opponents. In such a model it is interesting to investigate the effect of changing the value of k, which controls the amount of positive (negative) reinforcement of the individual strength after winning (losing) an interaction.

To show the influence of k, we made an artificial data set as in example II in which 13 animals meet each other in one distinct sequential order but in this case without deciding beforehand who wins or loses. Instead, the outcome of each subsequent dominance interaction was probabilistically determined by the difference in the current Elo-ratings of the contestants according to the table of win chances presented in the Appendix. That is, we assumed (for this simulation) that the actual dominance strengths of the individuals are given by the Elo-ratings calculated up to that moment, and these strengths determined the chances of winning and losing in the subsequent contests. At the start of each simulation run all individuals were given an initial rating of 1000. The simulations were run with three different values for k (20, 100, 200). The larger k is, the more leverage the winner of a previous interaction will have in subsequent encounters, since its rating increases more with a larger k value. Table 1 gives the resulting ordinations/ratings. In this simulation, in which the outcome of each interaction depends probabilistically on the difference in ratings of the contestants, early wins determine to a large extent the final ranking, especially when k is large; therefore the rankings in Table 1 all differ from each other.

Changing k has a clear influence on the ordination results; the larger k is, the stronger the influence of a win (or loss) on future interactions, just as a 'winners-bonus' is supposed to. This automatically results in a wider range of Elo-ratings when k is large.

When we look at the structural features in the resulting dominance matrices, no large differences are found for the different values of k. The dominance matrices turn out to be unidirectional for all choices of k. They are also significantly linear (Landau's $h \geqslant 0.99$ and $P<0.0001$ in all three cases). When the I&SI method is applied to the resulting dominance matrices the rankings found by this method turn out to be rather similar to the Elo-rating rank orders (Spearman's rho>0.85 and $P<0.001$ in all three cases). Of course, the rankings yielded by the I&SI method are not expected to be identical to the Elo-rating rank orders, since the Elo-rating method takes the sequential order in which the outcomes of the interactions are realized into account, whereas the I&SI method does not. Note that in contrast to example II above, a linear hierarchy has not been put into the data set beforehand, since the outcome of each dominance interaction has been made dependent here on the current difference in ratings between the opponents. So the establishment of a linear hierarchy is an emergent feature. Inspection of the development of the linearity of the dominance relationships over time (as measured by Landau's linearity index h) reveals that for a large k a stable linear rank order is soon established (Landau's h>0.9), whereas for a smaller value of k it takes longer for a linear hierarchy to become established.

We repeated the k=200 simulation 20 times to make a figure that can be compared to the results of the

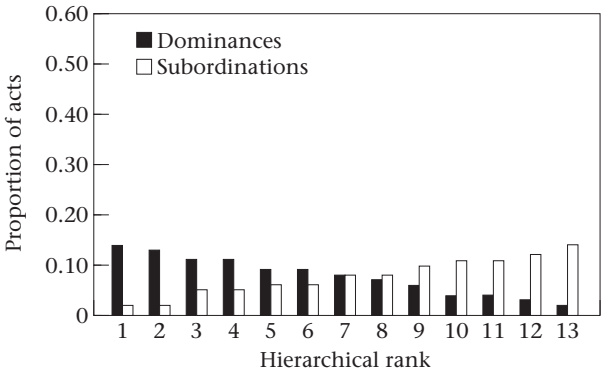

Figure 2. Distribution of dominances and subordinations (proportion of the total number of acts) as a function of hierarchical rank. Twenty data sets of 13 individuals were simulated and averages are presented. The *k* factor was set to 200. Simulation runs using $k=20$ and $k=100$ resulted in similar graphs.

simulation model of Theraulaz et al. (1995). In their model differences in resource-holding power are used in a similar way as in the Elo-rating method to determine the outcome of future agonistic encounters. In our study the number of interactions between any two individuals was random and about equal (3000) for all dyads. Figure 2 shows the distribution of dominances (wins) and subordinations (losses) as a function of hierarchical rank. The simulation based on the Elo-rating method yields an identical distribution as found by Theraulaz et al. in their simulations in which the probability of performing a dominance interaction when two individuals meet is kept constant (compare Fig. 2 with the topmost panels in Figure 4 of Theraulaz et al. 1995).

Discussion

The Elo-rating method estimates sequentially the dominance strengths of individuals on the basis of the outcomes (wins, losses, ties) of the dyadic agonistic dominance interactions and is as such a useful complement to the existing methods of dominance rank ordination. It is difficult to assess whether this Elo-rating could somehow be taken to represent the psychophysiological characteristics of the different individuals in any animal species. No obvious characteristic springs to mind other than resource-holding power (Pusey & Packer 1998). If resource-holding power comprises a combination of health, physical strength and mental capacities, to a certain extent it may also be assessed by the individuals themselves by observing their conspecifics.

When using Elo-rating as a model for generating dominance in a group we determined the outcomes of future dominance interactions by using the Elo-ratings of the subjects up to that moment as if these ratings are identical to the actual strengths. In this simulation the distribution of dominances and subordinations in relation to the established rank resembles the distribution obtained with the model developed by Theraulaz et al. (1995). In this model the chance of winning (P_{ij}) is determined by a so-called Fermi function of the difference in dominance strengths of the two contestants, i and j, $P_{ij}=1/(1+\exp(-\eta(D_i-D_j))$, where η is a temperature-like coefficient, and D_i is the dominance strength of individual i. This functional dependency is qualitatively similar to the one that is used in the Elo-rating method. Theraulaz et al.'s method of updating dominance strengths, however, is different, as dominance strengths are updated either by increasing or decreasing the current value with a constant, or by calculating the dominance strength, after each contest, as the ratio of number of won contests divided by the number of all contests involving individual i.

Boyd & Silk (1983) showed how the Bradley–Terry model can be used to estimate cardinal dominance ranks from a dominance interaction matrix. In this model the functional dependency between the probability of winning and the difference in dominance strengths is defined in the same way as in the Theraulaz (1995) model (apart from the temparature-like coefficient in this last model). Starting from an observed dominance interaction matrix and assuming this specific relationship between the difference in dominance strength and the probability of winning, the individual dominance strengths are estimated. Although the method of Boyd & Silk has the advantage over most dominance rank methods that their cardinal index of dominance rank is at the level of an interval scale (which also holds for the Elo-rating method), it disregards the sequential information, as all methods do that start from a data matrix instead of a data sequence. Indeed, when we apply Boyd & Silk's method for estimating cardinal dominance ranks to the matrix of the 33 dominance interactions shown in Fig. 1, we obtain cardinal ranks that are rather different from the Elo-ratings. Specifically, b and c both have high cardinal ranks (19.2 and 18.5 respectively), while d as the third ranking animal has only 9.5. Individuals f and g turn out to have the same cardinal rank of 0.0, while e obtains a value of 0.3 and animal a has the lowest value of -1.5. Differences between these cardinal ranks and the Elo-ratings (apart from the irrelevant scale difference of course; either scale can be linearly transformed) are due to the Boyd & Silk method taking the outcomes of all interactions into account equally, whereas in the Elo-rating method unexpected outcomes have a stronger influence on the Elo-ratings than outcomes that do not deviate from what is expected on the basis of the current ratings of the contestants (see the comparison of the 2nd with the 24th interaction both involving individuals c and g, above). Another difference between the methods is that Boyd & Silk's method often cannot be applied when the dominance matrix contains very few entries below the diagonal, owing to nonconvergence of the iterative estimation process (Boyd & Silk 1983, page 49), whereas for Elo-rating this presents no particular problem.

Another simulation model of dominance interactions, in this case among bumblebees, which also shows in part similar features to the Elo-rating method, has been developed by Hogeweg & Hesper (1983). In this model, dominance interactions are simulated by means of a

rule that specifies the upgrading after each interaction and also how the chance of winning depends on the strengths of the contestants. With respect to the upgrading this DoDOM rule equals the Elo-rating method: dominance strength upgrading is directly based on the difference in strength between the contestants and the deviation of the expected outcome from the actual outcome of the contest. However, the dependency of the chance of winning on the dominance strengths of the two interactions differs from both the Elo-rating method and the Theraulaz model. Specifically, when individuals i and j have a dominance interaction, the DoDOM rule specifies the chance of a win by i as being equal to the dominance strength of i divided by the sum of the strengths of i and j, that is, $P_{ij}=D_i/(D_i+D_j)$. So, in this model the chance of a win depends on the ratio of the dominance strengths between the two contestants rather than on the difference in strength between them.

As yet we do not see a limitation on the use of Elo-rating to determine a rank order on the basis of a sequence of observed dominance interactions. Obviously, a certain initial number of interactions is needed but this is fairly small and it is by no means necessary to have observations on all dyads in a matrix. In fact, if every individual has interactions with only two others of which one is above and one is below it in rank (except for the individuals highest and lowest in rank), this should already be sufficient. Obviously, the more interactions are observed the more reliable the Elo-ratings will agree with the actual strengths. As the data set is treated as a sequence instead of a matrix, one can always choose objectively some threshold beyond which the rank is treated as reliable.

We have already pointed out that for specific data sets a correlation between the ranking based on Elo-rating and the ranking obtained by a rank ordination method such as the I&SI method can be high and significant. Theoretically, such a correlation is not necessary, but in practice, there is some reason for concern if the Elo-ordination and conventional matrix-based ordinations are not to a certain extent similar, as this would mean that rank ordination is so heavily dependent on the actual sequence of events (and thus changes so fast) that rank as a variable might not be useful under such conditions. Differences between the rank orders obtained by the two methods give some indication of the level of rank stability; the more they are alike, the more stable the system is.

One remaining point of discussion is the choice of the constant k, and its consequences for the eventual outcome, in particular when dominance interactions can be won by means of different behaviours showing different intensities of aggression. Choice of k influences only the speed with which an individual's estimated rating rises or falls, in particular early in the sequence of observed interactions when estimated ratings still deviate strongly from actual ones. When dealing with real data sets an accurate choice of k will lead to an ordination that is reliable after a short observation period; choosing k too high or too low will not in the long run influence the resulting rank order, but it will take longer for the estimated ratings to approach the actual ones. Example I gives us a good reason to start low and eventually increase k for certain types of behaviour based on observation. If it turns out that certain behaviours such as biting occur only if the difference in rating is far larger than k, four times as large in example I, there is good reason to assume that for biting k is larger than the initial 20 and is maybe as high as 80. Note that this automatically mimics the effect of a winner's bonus for an opponent that dares to bite which is in line with the observations of Chase (1982, 1985, 1987). In fact, k can also be deduced from observations such as those of Chase. Chase (1982) determined, for instance, the chance that A will win from C, after A has won from B: this was 74%. This corresponds with an Elo-rating difference of 184 between A and C. Starting with all three individuals equal, then A could have achieved such rating difference only by winning from B who was until then estimated equal to A and thus both had even chances to win. Hence k must have been twice the established rating difference between A and C: 368. In the same experiment Chase showed that the chance of C winning from A after A has won from B is 4% (which corresponds with a k of 250) and is equal to the chance of B winning from C after A has won from B. That these chances are equal can easily be seen when using Elo-rating: if A wins from B, A's initial value increases with $0.5k$ and B's initial value decreases with $0.5k$. So the rating difference between A and C (C still at its initial value) is equal to the rating difference between B and C , which corresponds to the equal chances for 'C to win from A' and 'B to win from C'.

As Elo-rating continually estimates the individuals' strengths (or resource-holding potential) after each particular interaction, the question remains whether it keeps up with the temporal changes in these individual strengths. This obviously depends on the capabilities of the species involved, as well as on the choice of k, and whether a stable hierarchy has settled yet. Probably Elo-rating estimates will not trail far behind the actual dominance strengths, but we can imagine a species in which crucial dominance deciding events are sparse and might easily be missed in data sampling. One can question whether any other method would do better in such a case. If something like this is to be expected, Elo-rating might trail behind, but will probably still provide a better estimate at the end of the sequence than a method which requires the stability of the rank order for the whole observation period.

We thank A. L. Louwerse and K. Berkhout for allowing us to use their data. We also thank Peter Rothery and an anonymous referee for their useful comments on the manuscript.

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.

Bossuyt, P. 1990. *A Comparison of Probalistic Unfolding Theories for Paired Comparisons Data*. Berlin: Springer Verlag.

Boyd, R. & Silk, J. B. 1983. Method for assigning cardinal dominance ranks. *Animal Behaviour*, **31,** 45–58.

Brown, J. L. 1975. *The Evolution of Behavior*. New York: Norton.

Chase, I. D. 1982. Dynamics of hierarchy formation: the sequential development of dominance relationships. *Behaviour*, **80,** 218–40.

Chase, I. D. 1985. The sequential analysis of aggressive acts during hierarchy formation: an application of the 'jigsaw puzzle' approach. *Animal Behaviour*, **33,** 86–100.

Chase, I. D. & Rohwer, S. 1987. Two methods for quantifying the development of dominance hierarchies in large groups with applications to Harris' sparrows. *Animal Behaviour*, **35,** 1113–1128.

Chase, I. D., Bartolomeo, C. & Dugatkin, L. A. 1994. Aggressive interactions and inter-contest interval: how long do winners keep winning. *Animal Behaviour*, **48,** 393–400.

Clutton-Brock, T. H., Albon, S. D. & Guinness, F. E. 1979. The logical stag: adaptive aspects of fighting in red deer (*Cervus elaphus* L). *Animal Behaviour*, **27,** 211–225.

Crow, E. L. 1990. Ranking paired contestants. *Communications in Statistics: Simulation and Computation*, **19,** 749–769.

David, H. A. 1987. Ranking from unbalanced paired-comparison data. *Biometrika*, **74,** 432–436.

David, H. A. 1988. *The Method of Paired Comparisons*. New York: Hafner.

Elo, A. E. 1961. The new U.S.C.F. rating system. *Chess Life*, **16,** 160–161.

Elo, A. E. 1978. *The Rating of Chess Players, Past and Present*. New York: Arco.

Hogeweg, P. & Hesper, B. 1983. The ontogeny of the interaction structure in bumble bee colonies: a MIRROR model. *Behavioral Ecology and Sociobiology*, **12,** 271–283.

Jameson, K. A., Appleby, M. C. & Freeman, L. C. 1999. Finding an appropriate order for a hierarchy based on probabilistic dominance. *Animal Behaviour*, **57,** 991–998.

McMahan, C. A. & Morris, M. D. 1984. Application of maximum likelihood paired comparison ranking to estimation of a linear dominance ranking in animal societies. *Animal Behaviour*, **32,** 374–378.

Pusey, A. E. & Packer, C. 1997. The ecology of relationships. In: *Behavioral Ecology* (Ed. by J. R. Krebs & N. B. Davies), pp. 254–283. Boston: Blackwell.

Slater, P. 1961. Inconsistencies in a schedule of paired comparisons. *Biometrika*, **48,** 303–312.

Theraulaz, G., Bonabeau, E. & Deneubourg, J. L. 1995. Self-organization of hierarchies in animal societies: the case of the primatively eusocial wasp *Polistes dominulus* Christ. *Journal of theoretical Biology*, **174,** 313–323.

de Vries, H. 1998. Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. *Animal Behaviour*, **55,** 827–843.

de Vries, H. & Appleby, M. C. 2000. Finding an appropriate order for a hierarchy: a comparison of the I&SI and the BBS methods. *Animal Behaviour*, **59,** 239–245.

Appendix

Difference in Elo-rating and its corresponding expected chance of winning

Rating difference	Expected chance of winning	Difference	Chance	Difference	Chance
0>=dif<=3	0.50	122>=dif<=129	0.67	279>=dif<=290	0.84
4>=dif<=10	0.51	130>=dif<=137	0.68	291>=dif<=302	0.85
11>=dif<=17	0.52	138>=dif<=145	0.69	303>=dif<=315	0.86
18>=dif<=25	0.53	146>=dif<=153	0.70	316>=dif<=328	0.87
26>=dif<=32	0.54	154>=dif<=162	0.71	329>=dif<=344	0.88
33>=dif<=39	0.55	163>=dif<=170	0.72	345>=dif<=357	0.89
40>=dif<=46	0.56	171>=dif<=179	0.73	358>=dif<=374	0.90
47>=dif<=53	0.57	180>=dif<=188	0.74	375>=dif<=391	0.91
54>=dif<=61	0.58	189>=dif<=197	0.75	392>=dif<=411	0.92
62>=dif<=68	0.59	198>=dif<=206	0.76	412>=dif<=432	0.93
69>=dif<=76	0.60	207>=dif<=215	0.77	433>=dif<=456	0.94
77>=dif<=83	0.61	216>=dif<=225	0.78	457>=dif<=484	0.95
84>=dif<=91	0.62	226>=dif<=235	0.79	485>=dif<=517	0.96
92>=dif<=98	0.63	236>=dif<=245	0.80	518>=dif<=559	0.97
99>=dif<=106	0.64	246>=dif<=256	0.81	560>=dif<=619	0.98
107>=dif<=113	0.65	257>=dif<=267	0.82	620>=dif<=735	0.99
114>=dif<=121	0.66	268>=dif<=278	0.83	dif>=736	1.00

To calculate Elo-ratings after each contest this table is used, which gives for each rating difference between the two contestants the corresponding expected chance of winning for the one with the highest rating. Graphically this table resembles a logistic curve.