Neumann, C, Duboscq, J, Dubuc, C, Ginting, A, Irwan, AM, Agil, M, Widdig, A and Engelhardt, A

Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating

Neumann, C, Duboscq, J, Dubuc, C, Ginting, A, Irwan, AM, Agil, M, Widdig, A and Engelhardt, A (2011) Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating. Animal Behaviour, 82 (4). pp. 911-921. ISSN 0003-3472

Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo rating

Keywords: Elo rating, dominance rank, dominance hierarchy, methodology, *Macaca nigra*, *Macaca mulatta*, I&SI, David’s score

Dominance is one of the most important concepts in the study of animal social behaviour. Dominance hierarchies in groups arise from dyadic relationships between dominant and subordinate individuals present in a social group (Drews 1993). High hierarchical rank or social status is often associated with fitness benefits for individuals (e.g., Côté & Festa-Bianchet 2001; von Holst et al. 2002; Widdig et al. 2004; Engelhardt et al. 2006), and hierarchies can be found in most animal taxa including insects (e.g., Kolmer & Heinze 2000), birds (e.g., Kurvers et al. 2009) and mammals (e.g., Keiper & Receveur 1992).

The analysis of dominance has a long-standing history (Schjelderup-Ebbe 1922; Landau 1951), and a great number of methods to assess hierarchies in animal societies are currently available (reviewed in de Vries 1998; Bayly et al. 2006; Whitehead 2008). Though differing in calculation complexity, all ranking methods presently used in studies of behavioural ecology are based on interaction matrices. For this, a specific type of behaviour or interaction, from which the dominance/subordinance relationship of a given dyad can be deduced, is tabulated across all individuals (see for example, Vervaecke et al. 2007). This matrix can either be reorganized as a whole in order to optimize a numerical criterion (e.g., I&SI: de Vries 1998; minimizing entries below the matrix diagonal: Martin & Bateson 1993), or alternatively, an individual measure of success calculated for each animal present (e.g., David's score: David 1987; CBI: Clutton-Brock et al. 1979). In the latter case, a ranking can be generated by ordering the obtained individual scores.

Although calculations of dominance hierarchies are routinely undertaken in many studies of behavioural ecology, and although there have been numerous methodological developments in this area (e.g. Clutton-Brock et al. 1979; David 1987; de Vries 1998), there are still a number of obstacles and limitations scientists have to tackle when analysing dominance relationships. This is mainly due to the fact that the methods commonly used can often not be applied to highly dynamic animal societies, or to sparse data sets, and because methods based on interaction matrices need to fulfil certain criteria in order to generate reliable results. Generally, many researchers may not be aware of some of the problems that are associated with the application of such methods to their data sets, which may in the worst case lead to the misinterpretation of results.

An alternative method that can overcome the shortcomings of matrix-based methods is Elo rating. Developed by and named after Arpad Elo (Elo 1978), it is used for ratings in chess and other sports (e.g., Hvattum & Arntzen 2010), but has been rarely used in behavioural ecology (but see Rusu & Krackow 2004; Pörschmann et al. 2010). The major difference to commonly used ranking methods is that Elo rating is based on the sequence in which interactions occur, and continuously updates ratings by looking at interactions sequentially. As a consequence, there is no need to build up complete interaction matrices and to restrict analysis to defined time periods. Ratings (after a given start-up time) can be obtained at any point in time, thus allowing monitoring of dominance ranks on the desired time scale.

The major aim of this paper is to promote Elo rating amongst behavioural ecologists by illustrating its advantages over common methods, and by validating its reliability for assessing dominance rank orders, particularly in highly dynamic social systems. By providing the necessary computational tools along with an example (see electronic supplementary materials), we also make Elo rating user-friendly. In the following, we start with an introduction into the procedures of Elo rating. We then show that with Elo rating it is easy to track changes in social hierarchies, which may be overlooked with matrix based methods, and point out several general advantages of Elo rating over matrix based methods. In order to demonstrate the benefits of Elo rating empirically, we present the results of a reanalysis of one of our own previously published datasets. Finally, we validate the reliability and robustness of Elo rating by comparing the performance of this method with those of two currently widely used ranking methods, the I&SI method and the David's score, using empirical data and reduced data sets that mimic sparse data.

Elo Rating Procedure

Elo rating, in contrast to commonly used methods, is not based on an interaction matrix, but on the sequence in which interactions occur. At the beginning of the rating process, each individual starts with a predefined rating, for example a value of 1000. The amount chosen here has no effect on the differences in ratings later: the relative distances between individual ratings will remain identical (Albers & de Vries 2001). After each interaction, the ratings of the two participants are updated according to the outcome of the interaction: the winner gains points, the loser loses points. The amount of points gained and lost during one interaction depends on the expectation of the outcome (i.e., the probability that the higher rated individual wins, Elo 1978) prior to this interaction. Expected outcomes lead to smaller changes in ratings than unexpected outcomes (Figure 1). Depending on whether the higher rated individual wins or loses an interaction, ratings are updated according to the following formulae:

Higher rated individual wins:

Eq1: $\text{WinnerRating}_{new} = \text{WinnerRating}_{old} + (1 - p) \times k$

Eq2: $\text{LoserRating}_{new} = \text{LoserRating}_{old} - (1 - p) \times k$

Lower rated individual wins (against the expectation):

Eq3: $\text{WinnerRating}_{new} = \text{WinnerRating}_{old} + p \times k$

Eq4: $\text{LoserRating}_{new} = \text{LoserRating}_{old} - p \times k$

where p is the expectation of winning for the higher rated individual, which is a function of the absolute difference in the ratings of the two interaction partners before the interaction (Figure 1; see also Elo 1978; Albers & de Vries 2001). k is a constant and determines the amount of rating points that an individual gains or loses after a single encounter. Its value is usually set between 16 and 200 and once chosen remains at this value throughout the rating process. In the short term, k influences the speed with which Elo ratings increase or decrease. In the long term, however, k appears to have only minor influence on the rankings obtained (Albers & de Vries 2001, Neumann et al. unpubl. data). For the latter reason, we used an arbitrary fixed $k = 100$ throughout our analyses, even though the choice of k can have interesting implications (see section Integrity of Power Assessment).

As Elo rating estimates competitive abilities by continuously updating an individual's success, it reflects a cardinal score of success. As such, the differences between ratings are on an interval scale and may thus allow the application of parametric statistics in further analyses. An example, illustrating the process of Elo rating in more detail, can be found in appendix 1 (see also Albers & de Vries 2001).

Advantages of Elo Rating over Matrix Based Methods

No minimum number of individuals

Scientists often face the problem of small sample sizes when it comes to determining dominance hierarchies. In many group living species, age-sex classes or even complete groups contain less than six individuals. Problems with matrix-based methods therefore start with the calculation of linearity (i.e., if A is dominant over B and B is dominant over C, then A is dominant over C). The commonly used index to assess the degree and statistical significance of linearity (Landau 1951; de Vries 1995), will only yield significant results if the number of individuals in the matrix exceeds five individuals (Appleby 1983), thus preventing, for example, the application of the widely used I&SI method (de Vries 1998) to small groups.

Elo rating, however, can be applied to groups of any size with only two individuals required for the calculation of Elo ratings (see Figure 1).

Independence of Demographic Changes

Biological systems are often very dynamic in regard to group composition. New offspring is born, maturing animals migrate, individuals become the victim of predation, floating individuals may join groups temporarily, or entire groups fission and fusion regularly.

An advantage of Elo rating is the incorporation of demographic changes such as migration events without interruption of the rating process itself. Whereas matrix based methods need to discontinue rating and to build up new matrices (which then need a sufficient number of interactions between individuals in order to produce reliable rankings) after each demographic change, hierarchy determination can be continued despite demographic changes. This is achieved by giving a new individual the predefined starting value (as defined for all individuals before they are rated for the first time) before the first interaction with another individual. After a few interactions this individual can be ranked in the existing hierarchy (see below). This feature may be particularly advantageous for studies on species that live in large social groups with high reproductive rate, high migration rate and/or high predation rate.

To illustrate this, we plotted the development of Elo ratings of adult males in a group of crested macaques over the course of a month during which three migration events took place (Figure 2, see below for details on the study population and data collection). In our example, male ZJ migrated into group R2 on March 11th, 2007. To include him in the dominance hierarchy, he was assigned the initial score of 1000, and even though he lost his first observed interaction, Elo rating made it possible to recognize him quickly as the new alpha male. Likewise, individuals that emigrate (or die) (like males SJ and YJ in this example) are simply excluded from the rating process from the date of their disappearance without causing any interruption to the rating procedure.

Since Elo rating does not stop the rating process as a consequence of changes in group composition it circumvents a further drawback of matrix-based methods. Techniques such as I&SI and David's score result in values that directly depend on the number of individuals present, thus an observed change in calculated dominance rank or score across two time periods may in fact be a consequence of changes in the number of animals in the group rather than changes in competitive abilities, thus making a comparison invalid. For example, in the case of the normalized David's score (c.f. de Vries et al. 2006), values can range between 0 and $N - 1$, where N is the number of individuals present in the social group. Elo rating, in contrast, results in ratings that do not depend on the number of individuals present. Given that k is fixed for the entire rating process, the current opponent's strength is the only variable that influences an individual's future rating. Hence, the Elo rating of an individual is independent of the number of individuals, and time periods that need to be created as a consequence of changes in the number of individuals. This feature allows Elo rating to be used in a longitudinal manner which is crucial for a wide array of studies, e.g., those on mechanisms of rank acquisition and maintenance, determinants of life-time reproductive success, and so on.

However, as in the other methods, true ratings of individuals are only known after a minimum amount of interactions involving these individuals occurred (see also Albers and de Vries 2001). For example (Figure 2), rank orders that would have been obtained through Elo rating within the first two weeks of ZJ's group membership would have placed him as ranking below BJ. After 13 days (i.e., eight observed interactions), ZJ reached the top-ranked position in the Elo ratings. Using all observed interactions from these two weeks it was not possible to construct a linear hierarchy, and only after 45 days did we obtain a matrix with a sufficient amount of interactions permitting the use of I&SI. However, it is likely that ZJ became alpha male directly upon his arrival in the group even though he lost his very first observed interaction (top entry: see e.g., Sprague et al. 1998) rather than constantly rising through the hierarchy. Albers and de Vries (2001) suggest waiting for at least two interactions before assessing a dominance hierarchy through Elo rating whenever a new member joins the hierarchy: one against a stronger and one against a weaker opponent. In the case of ZJ, however, we observed him interacting with six out of the seven other males present. In our case it thus seems more appropriate to follow Glickman and Doan's (2010, rating chess players) suggestion to treat ratings based on less than nine interactions as 'provisional' and exclude such ratings from rankings. Therefore in general, Elo rating still needs a short start-up time before creating reliable dominance hierarchies when group composition changes. This start-up time is however much shorter than the time needed to build up sufficiently filled interaction matrices for dominance hierarchies.

Visualization and Monitoring of Hierarchy Dynamics

Even if group composition is stable, matrices do not allow dynamics to be tracked within social hierarchies, especially if study periods are very short and data insufficient to obtain reliable rankings. In the worst case, a researcher may overlook rank changes when analysing hierarchies at some fixed interval (e.g., monthly).

One of the great advantages of Elo rating is its ability to visualise dominance relationships on a time scale, thus allowing monitoring of rank relationship dynamics. As the information about the sequence of interactions is a prerequisite for applying Elo rating, one can easily create graphs that depict the time scale on the x-axis and plot the development of each individual's ratings on the y-axis. This approach can demonstrate a fundamental feature of Elo rating, i.e., the possibility to obtain a rank order at any given point in time by ordering the most recently updated ratings for a given set of individuals. For example (Figure 2), the ordinal rank order among the present individuals on March 1st based on Elo ratings was SJ (1810 Elo points), BJ (1592), YJ (1317), VJ (1068), KJ (982), TJ (942), RJ (703), CJ (526), PJ (90). By March 31st, however, the ordinal rank order had changed into ZJ (1355), BJ (1262), VJ (994), TJ (950), KJ (892), RJ (600), CJ (592), PJ (53).

Figure 3 gives an example illustrating how Elo rating can reflect dynamics in rank relationships. In late June 2007, medium ranked male KJ started losing interactions against several lower ranked males and dropped to rank eleven. As such, his drop to the lowest rank among group males is reflected by a quick decrease in his Elo rating by several hundred points in only a few days (Figure 3). Such dynamics are difficult to track with both I&SI and David's score since a new matrix would need to be created after such a conspicuous event, requiring a sufficient amount of data to obtain reliable rankings.

At the same time, it is common practice to calculate dominance hierarchies based on rather arbitrary time period definitions (e.g., monthly: Silk 1993; Setchell et al. 2008). This might lead to blurring or in the most extreme case even to overlooking dynamics in rank relationships. Elo rating, with its capacity to visualize dominance relationships graphically, allows identification of such dynamics in rank relationships in great detail. Hierarchies for the example month June 2007 (Figure 3) obtained with matrix based methods lead to illogical rankings: the I&SI algorithm assigns KJ rank 11, whereas David's score ranks KJ 10^{th} (note that linearity is statistically significant during this month: $h' = 0.50$, $P = 0.043$, total of 205 interactions, 24% unknown relationships). Elo rating, in contrast, shows that KJ held a medium rank almost throughout the entire month and dropped in rank only during the last week of June.

In Old World monkeys and many other group living mammals, it is sometimes observed that young males rise in rank before they eventually leave their natal group (e.g., Hamilton & Bulger 1990). A common approach to quantify this phenomenon would be to calculate monthly ranks and correlate them with the time to departure. Doing so for 16 natal male crested macaques (see below for details on the study population and data collection) using David's score, however, lends only little support to this phenomenon (Spearman's rank correlation: $r_s = 0.642$, $P = 0.139$, $N = 7$, Figure 4a). As described below, this may be the consequence of high proportions of unknown relationships leading to less reliable scores. It could also be due to the fact that David's scores directly depend on the number of individuals incorporated in the matrix. In contrast, when using Elo rating, the hypothesis that natal males rise in rank before emigration is strongly supported ($r_s = 1$, $P < 0.001$, $N = 7$, Figure 4b). We observe an almost linear increase in ratings before the migration date. It appears that males went through a noticeable surge about three months before emigration, and kept rising before their departure. This is, however, a preliminary result and further investigation is warranted. Since Elo ratings can be obtained at any desired date, even an analysis with higher time resolution (e.g., weekly) is possible (Figure 4c).

In addition, Elo rating also allows objective identification and quantitative characterization of hierarchical stability. Again, the graphical features of Elo rating provide very useful assistance in this respect. Figure 2, for example, shows that individuals KJ and TJ changed their ordinal rank relative to each other five times within one month, suggesting some degree of rank instability (see also individuals RJ, TJ and GM in Figure 3).

To quantify the degree of hierarchy stability, we propose to use the ratio of rank changes per individuals present over a given time period. Formally, the index is expressed as

Eq5: $$S = \frac{\sum_{i=1}^{n} (C_i \times w_i)}{\sum_{i=1}^{n} N_i},$$

where C_i is the sum of absolute differences between rankings of two consecutive days, w_i is a weighing factor determined as the standardized Elo rating of the highest ranked individual involved in a rank change, and N_i is the number of individuals present on both days (see appendix 2 for further details). Before division, values are summed over the desired time period, i.e. *n* days. *S* can take values between 0, indicating a stable hierarchy with identical rankings on each day of the analyzed time period , and 2 / max(N_i), indicating that the hierarchy is reversing every other day, i.e. total instability. Our data suggest that *S* typically ranges between 0 and 0.5.

To test the validity of this approach we calculated *S* before and after the immigration of male macaques that subsequently achieved high ranks (among the top three, see below for details on the study population and data collection). We expected such events to induce instability (e.g., Lange & Leimar 2004; Beehner et al. 2005), thus leading to higher *S* values when compared to periods before such incidents. We found less stability, i.e. greater *S* values, during four-week periods after the immigration of males that achieved high rank compared to the four-week periods before (Wilcoxon signed rank test: $V = 87$, $N = 14$, $P = 0.030$), indicating that hierarchies were less stable after the immigration of a high ranking male. In contrast, after the immigration of males that subsequently held low ranks, we observed no such difference in stability ($V = 14$, $N = 7$, $P = 1.000$).

Such a quantitative approach may be advantageous since, so far, hierarchical instability has been identified in a non-consistent manner. Sapolsky (1983) for example, studying baboons, identified periods of instability in male dominance hierarchies through high rates of ambiguously ending agonistic interactions and through high rates of interactions that ended with the subordinate winning. In a different study of baboons, Engh et al (2006) assessed instability in female dominance hierarchies in a mere descriptive way. On a long-term basis, stability has also been characterised by comparison of rankings in consecutive seasons using regression or correlation analysis (e.g., in mountain goats, Côté 2000). By objectively defining stability, Elo rating may become an important tool for studies on social instability and its consequences, for example on individual stress levels and health (e.g., Sapolsky 2005), territory acquisition (e.g., Beletsky 1992) or group transfer (e.g., Smith 1987; van Noordwijk & van Schaik 2001). In addition, the objective quantification of stability may make comparisons across studies possible.

Independence of Time Periods

It is common practice to obtain hierarchies at some arbitrary fixed time interval (e.g. monthly). Given the dynamics of animal societies, both in group composition and rankings (see above), such an approach is prone to misjudgement of hierarchies for two reasons. First, all individuals incorporated in a dominance matrix must have the possibility to interact with each other at all times. If group composition changes within the studied interval, for example in fission/fusion societies or when individuals leave and join frequently (floaters), applying matrix based methods is unjustified. Second, rank changes that occur will be blurred (see the example above, Figure 3).

With Elo rating it is possible to pinpoint rankings to a specific day. This is of particular importance when studying events, such as a male's rank at the day his offspring was conceived or born, or tracking the rank development of individuals before and after they migrate.

A related problem to the creation of time periods is the proportion of unknown relationships. When creating relatively short time periods to account for the above mentioned dynamics, one often faces a high percentage of pairs of individuals that were not observed interacting in a given period. Like any statistical test, ranking methods suffer from decreased power or precision when sample size is low (Appleby 1983; de Vries 1995; Koenig & Borries 2006; Wittemyer & Getz 2006), even though attempts have been made to counter this problem (see de Vries 1995, 1998; de Vries et al. 2006; Wittemyer & Getz 2006).

As we will show below, Elo rating seems less affected by unknown relationships than matrix based methods, and is therefore also operational on very sparse data sets.

Integrity of Power Assessment

Without demonstrating their application, we finally mention three further advantages of Elo rating that may refine the precision of power assessment of individuals: a) integration of undecided interactions into the rating process, b) discrimination of agonistic interactions of differing quality, and c) choosing *k* according to the study species.

Undecided interactions

Though some matrix-based methods (e.g., David's score or Boyd and Silk's (1983) index) explicitly allow interactions without unambiguous winners and losers, i.e., draws or ties, to be taken into account when establishing dominance orders, researchers (including us) usually choose to discard such observations. Clearly, agonistic interactions that end without unambiguous winners and losers contain information about competitive abilities of the involved individuals and should therefore not be disregarded. When using Elo rating, an undecided interaction can be incorporated into the rating process to the disadvantage of the higher rated individual whose rating will decrease, even though the decrease will be smaller than had the higher rated individual lost the interaction (Albers & de Vries 2001). After a draw the rating for the higher rated individual is reduced to $\text{Rating}_{new} = \text{Rating}_{old} - k\ (p - 0.5)$, whereas the rating for the lower rated individual increases to $\text{Rating}_{new} = \text{Rating}_{old} + k\ (p - 0.5)$. Hence, a draw between two individuals that had identical ratings before the interaction (i.e., $p = 0.5$) will not alter the ratings. In this way, Elo rating allows for a more complete power assessment of individuals by including interactions into the rating process that are just as meaningful as clear winner-loser interactions.

Agonistic interactions of different quality

Instead of being fixed throughout the rating process, the constant k could be adjusted according to the quality of the interaction or the experience of the interacting individuals. For example, one could distinguish between low- and high-intensity aggression (e.g., Adamo & Hoy 1995; Lu et al. 2008) and assign interactions involving high-intensity aggression higher values of k. This results in greater changes in ratings after such interactions compared to interactions involving low-intensity aggression.

Choosing k

Prior experience of individuals plays an important role in the outcome of agonistic encounters in many animal taxa: the winner of a previous interaction is more likely to win a future interaction, whereas losers are more likely to lose future interactions (Hsu et al. 2006). A meta-analysis on the magnitude of such winner/loser effects demonstrated that the likelihood of winning an interaction is almost doubled for previous winners whereas for previous losers the likelihood of winning is reduced almost five-fold (Rutte et al. 2006). Depending on the size of this effect in the study species, k could therefore be split into a smaller k_w for the winner and a larger k_l for the loser to reflect this phenomenon (de Vries 2009).

Thus, Elo rating is not limited to decided dominance interactions, but can incorporate undecided interaction and in addition allows for a detailed hierarchy evaluation by weighing interactions according to their properties and the magnitude of winner/loser effects. This surplus of information Elo rating can utilize allows for a much finer assessment of dominance relationships.

Testing the Reliability and Robustness of Elo Rating

So far, we have shown how Elo-rating circumvents the problems associated with matrix based methods. However, we have not yet shown how it compares to other methods in terms of reliability and robustness. We now compare Elo-rating with two widely used ranking methods that are based on interaction matrices (I&SI and David's score), using our own empirical data. Mimicking a variety of social systems, we use data collected on two species of macaques with different aggression patterns, crested (*Macaca nigra*, aggressive interactions frequent, but of low intensity) and rhesus macaques (*M. mulatta*, aggressive interactions less frequent, but of higher intensity) (de Waal & Luttrell 1989; Thierry 2007), and calculate dominance hierarchies for females (more stable hierarchies) and males (more dynamic hierarchies) separately. To facilitate the assessment of these analyses we will first briefly review the two methods we use for our comparisons.

Short Introduction to I&SI and David's Score

The I&SI method (de Vries 1998) is an iterative algorithm that tries to find the rank order that deviates least from a linear rank order. It is based on observed dominance interactions (e.g., winning/losing an agonistic interaction) and tries to minimize the number of inconsistencies (I) produced when building a dominance hierarchy, i.e., minimize dyads for which the relationship is not in agreement with the actual rank order. Subsequently, the strength of inconsistencies (SI), i.e., the rank difference between two individuals that form an inconsistency, is minimized, under the condition that in the iterated rank order the number of inconsistencies does not increase. The result of the I&SI algorithm is an ordinal rank order.

David's score (David 1987) is an individual measure of success, in which for each individual a score is calculated based on the outcome of its agonistic interactions with other members of the social group as $DS = w + w_2 - l - l_2$, where w is the sum of an individual's winning proportions and l the summed losing proportions. w_2 represents an individual's summed winning proportions (i.e., w) weighed by the w values of its interaction partners and likewise, l_2 equals an individual's summed losing proportions (i.e., l) weighed by the l values of its interaction partners (David 1987; Gammell et al. 2003; see de Vries et al. 2006 for an illustrative example). Thus, David's score takes the relative strength of opponents into account, valuing success against stronger individuals more than success against weaker individuals.

Rank orders generated with I&SI and David's score are generally very similar to each other (e.g., Vervaecke et al. 2007, Neumann et al. unpublished data).

Methods

Study populations

For our tests of Elo rating, we chose two species of macaques (crested, *Macaca nigra*, and rhesus macaques, *M. mulatta*). Even though our aim was not to test for species differences, we nevertheless aimed at gathering a broad data set including different, but comparable, species. Macaques fit this condition as the different species are characterised by a common social organization but at the same time by pronounced differences in aggression patterns (Thierry 2007).

Data collection

Between 2006 and 2010, we collected data in three groups (R1, R2, PB) of a population of wild crested macaques in the Tangkoko-Batuangus Nature Reserve, North Sulawesi, Indonesia (1º33' N, 125 º10' E; e.g., Duboscq et al. 2008; Neumann et al. 2010). Groups comprised between 4 – 18 adult males and 16 – 24 adult females and were completely habituated to human observers and individually recognizable. Between 2007 and 2010, data on rhesus macaques were collected in two groups (V, R) on the free ranging population on Cayo Santiago, Puerto Rico (18°09' N, 65°44' W). The study groups comprised between 20 – 60 females and 16 – 54 males (e.g., Dubuc et al. 2009, Widdig unpublished data).

We collected data on dyadic dominance interactions, i.e., agonistic interactions with unambiguous winner and loser, and displacement (approach / leave) interactions during all occurrence sampling on focal animals and during ad libitum sampling (Altmann 1974). Overall, our data set comprised a total of 12,740 interactions involving 252 individuals. Dominance hierarchies were created separately for the different species, groups and sexes.

Data analysis

Our first aim was to investigate whether dominance rank orders calculated with Elo rating reflect rankings obtained with more established methods. To answer this, we assessed how similar rank orders generated with Elo rating are to those obtained with the I&SI method and David's score. From our data on both macaque species, we created time periods based on socio-demographic events, such as changes between mating- and birth season, migration or death of individuals, maturing of subadult individuals and conspicuous status changes (hereafter "full data set", see Table 1) and produced corresponding dominance interaction matrices. Two consecutive time periods of a given species/sex combination did not comprise the same set of individuals in the majority of cases (61 out of 66 periods, i.e., 92%).

We tested all 66 matrices for linearity by means of de Vries' (1995) *h*' index. For the 29 matrices for which the linearity test yielded a significant result, we applied de Vries' (1998) I&SI method. Next, we calculated normalized David's scores from all matrices following de Vries et al. 2006. Finally, we calculated Elo ratings from all interactions in each of the group/sex combinations as a whole using Elo ratings on the last day of each time period for the comparison with I&SI ranks and David's scores. Elo ratings were calculated with 1000 as initial value and *k* was set to 100.

We computed Spearman's rank correlation coefficients between the rankings and scores for each period. To obtain positive correlation coefficients consistently for all comparisons, we reversed I&SI rank orders (i.e., high-ranking individuals get a high I&SI rank value), since high dominance rank is represented by high David's scores and Elo ratings. Thus, if two rankings are identical the correlation coefficient will be 1.00. We present average correlation coefficients with inter-quartile ranges. All calculations and tests were computed in R 2.12.0 and R 2.13.0 (R Development Core Team 2010). A script and manual to calculate and visualize Elo ratings with R along with an example data set can be found in the electronic supplementary material.

In a second analysis, we explored whether Elo rating is a robust method under conditions of sparse data and whether the performance of Elo rating under such conditions is systematically related to the percentage of unknown relationships in the interaction matrix. Please note that a sparse matrix is not necessarily a matrix with a higher proportion of unknown relationships. For example, a matrix in which each dyad was observed five times and all entries are above the diagonal (i.e., there are no unknown relationships) is more sparse than a matrix with each dyad being observed ten times (likewise, no unknown relationships). Whereas the I&SI ranking will be identical in both cases, David's scores will differ between the two, as will Elo ratings based on the interactions leading to this matrix.

We created sparse interaction matrices by randomly removing 50% of the observed interactions in each of the 66 time periods ("reduced data set": Table 1). These additional matrices were again tested for linearity, resulting in 17 matrices retaining significant linearity and thus justifying the application of the I&SI algorithm. We then calculated for each of the three methods separately correlation coefficients between rankings obtained from full and reduced data sets. For the 49 matrices that did not allow the use of I&SI due to non-significant linearity, we restricted the analysis to Elo rating and David's score.

To explore the robustness of the method further, we tested whether Elo rating is affected by increased proportions of unknown relationships and how it compared to the two other methods. In other words, we investigated whether the methods become less reliable as the proportion of unknown relationships increases. An increase in unknown relationships was generated as a consequence of the random deletion of 50% of all observed interactions (increase per period on average: 12.5%, inter-quartile range: 8 – 17%, "reduced data set": Table 1). We tested for an association between the increase in unknown relationships and the correlation coefficient between ratings from the full and reduced data set.

Results

Our results show that Elo ratings correlated highly with both I&SI ranks (median r_s = 0.97, quartiles: 0.94–0.99, N = 29 periods) and David's scores (median r_s = 0.97, quartiles: 0.96–0.99, N = 29 periods).

We found that Elo ratings from the full data set correlated highly with Elo ratings from the randomly reduced data set (Table 2). The performance of Elo rating is virtually identical to the one of I&SI and slightly higher compared to David's score (Table 2). Similarly, Elo rating produced strong correlations with slightly higher correlation coefficients compared to those obtained with David's score from the remaining 49 time periods for which I&SI could not be applied (Table 2).

Whereas there was no relationship between the increase in unknown relationships and the correlation coefficient between full and reduced data sets for Elo rating (r_s = –0.07, N = 17, P = 0.799) and I&SI (r_s = –0.36, N = 17, P = 0.162), we found that as the proportion of unknown relationships increased the correlation coefficients decreased between rankings from full and reduced data sets when using David's score (r_s = –0.52, N = 17, P = 0.031, Figure 5). Controlling for the initial proportion of unknown relationships by means of a partial Spearman correlation test leads to similar results (Elo rating: r_s = –0.02, N = 17, P = 0.927; I&SI: r_s = –0.39, N = 17, P = 0.110; David's score: r_s = –0.59, N = 17, P = 0.006),

Overall, our results indicate that Elo rating produces rank orders very similar to those obtained with I&SI and David's score. In addition, results of our tests suggest that rankings from Elo rating and I&SI (given significant linearity test) remain stable in sparse data sets, whereas David's score seems to create less reliable hierarchies in sparse data sets as a result of an increase in unknown relationships.

Discussion

Even though there is abundant literature available that compares the concordance of different methods for the assessment of dominance ranks or scores (e.g., Bayly et al. 2006; Bang et al. 2010), this is the first study to test the reliability of Elo rating with an extensive data set based on observations of free-ranging animals. Our results on dominance interactions in crested and rhesus macaques show that Elo rating produces dominance rank orders which closely resemble rankings generated with David's score and the I&SI method. Furthermore, our results indicate that Elo rating is very robust when data sets are limited in the number of interactions observed. Elo rating (and I&SI) even seems to produce more reliable dominance hierarchies than David's score when the proportion of unknown relationships is high. One could argue that this effect is due to the initial proportion of unknown relationships, i.e., a relatively high proportion of unknown relationships in a "full" matrix leads to some uncertainty in the ranking which may make the scores from the further reduced matrix even less reliable. However, when controlling for the initial proportion of unknown relationships, our results show that the robustness of Elo rating (and I&SI) is not attributable to this factor.

Using Elo Rating – an Example

We here demonstrate in an empirical example how Elo rating can improve study results due to its immunity to detrimental effects of assessing dominance status. Data for this example derives from a previous study where we investigated the relationship between dominance status and acoustic features of loud calls in male crested macaques (Neumann et al. 2010). We analyzed seven acoustic parameters and found three of them to be related to dominance status. However, due to frequent migration events and rank changes, and consequently short time periods with high percentages of unknown relationships, we were able to classify dominance only broadly into three rank categories (high, medium, low).

We reanalyzed our original data, using general linear mixed models (R package lme4: Bates et al. 2011, see Neumann et al. 2010 for details on the acoustic analysis and model specifications), and fitted separate models for each acoustic parameter, using Elo ratings from the day a loud call was recorded as predictor variable instead of rank categories. We additionally fitted models using monthly David's scores as predictor of dominance status.

In addition to the three parameters that we originally found to be affected by dominance rank, using Elo rating as predictor revealed two more acoustic parameters to be significant at $P < 0.05$ (corrected for multiple testing after Benjamini and Hochberg (1995), P values were assessed with the package languageR (Baayen 2011)). Using

Akaike's information criterion (AIC) to assess how well the models fitted the data (see, e.g., Johnson & Omland 2004), we found that of the five models yielding significant effects of Elo rating, four had smaller AIC values and thus fitted our data better than the respective models using rank categories as predictor. Surprisingly, when using David' scores as predictor, in none of the models did we find significant effects of dominance status after correction for multiple testing.

General Discussion

We have shown that Elo rating has several important advantages over common methods, such as the potential to: 1) monitor the dynamics of hierarchies and extract rank scores flexibly at any given point in time; 2) detect rank changes; 3) objectively identify hierarchy stability; 4) visualise hierarchy dynamics; 5) incorporate demographic changes into the rating procedure; 6) compare periods differing in demographic composition; 7) incorporate undecided interactions; and 8) objectively adjust the rating process based on species specific information.

We furthermore showed that Elo rating can increase power of analyses and explain more variation in our data under certain circumstances. Whether a reanalysis using Elo rating (as described above) will recover unexplained variation in general or not will mostly depend on how severe the potential negative effects of the data were on the ranks derived from matrices. For example, analysing a data set based on a single matrix with few unknown relationships will probably give very robust results, using either

David's Score or I&SI. Elo rating, in such a case will probably replicate the results obtained already, but not necessarily improve model fit. In contrast, a cross-sectional study on several groups, varying in the number of individuals and/or with high proportions of unknown relationships (as in our example above), may warrant a reanalysis using Elo rating.

We can however see one context in which Elo rating may not be the first choice to assess rank relationships. Unlike the I&SI method (given its application is feasible), Elo ratings do not necessarily reflect the rank order corresponding to a linear hierarchy in which an alpha individual is dominant (c.f., Drews 1993) over all other individuals and a beta individual is dominant over all other individuals except the alpha, and so on (de Vries 1998). Such a feature of a ranking algorithm may be desirable when, for example, investigating the relationship between parental and offspring rank (Dewsbury 1990; East et al. 2009; reviewed in Holekamp & Smale 1991). Such a situation is found in the matrilineal rank organization of many Old World monkeys, which is characterized by a linear structure in which a daughter ranks below her mother, and among all daughters of one mother the youngest one ranks highest (Kawamura 1958; Missakian 1972; but see Silk et al. 1981). Elo rating nevertheless produces rankings close to a linear hierarchy (see above), and may therefore still allow for appropriate rank assessment in such cases, especially when the I&SI method cannot be applied due to data limitations.

In conclusion, all the advantages mentioned in this paper make Elo rating a useful tool for assessing and monitoring changes of dominance relationships – particularly in highly dynamic animal systems.

Appendix 1

In this section, we give a detailed example of how Elo ratings are calculated. Figure and equation references refer to the main article.

To illustrate the principles of Elo rating, it is useful to consider the basic unit of any dominance hierarchy, the dyad. In the example presented here, two individuals A and B interact through a sequence of four interactions. At the start of this sequence their competitive abilities are unknown and thus there is no knowledge of their ratings, and both A and B are assigned an initial rating of 1000. At this stage of the rating process, both individuals are expected to be equally likely to win an interaction between each other since there is not yet a higher rated individual, i.e., $p = 0.5$. If A wins the first interaction against B, the ratings will be updated to $Elo_A = 1000 + (1 - 0.5) \times 100 = 1050$ (Eq1) and $Elo_B = 1000 - (1 - 0.5) \times 100 = 950$ (Eq2) (Figure 1: Interaction 1). Individual A thus gained 50 points whereas B lost 50 points. Given that A has won the first interaction, A is expected to win the next interaction against B with $p = 0.64$ due to the rating difference between A and B of 100 (Figure 1: Interaction 2, upper panel). If A wins the second interaction, ratings will be updated as follows: $Elo_A = 1050 + (1 - 0.64) \times 100 = 1086$ (Eq1) and $Elo_B = 950 - (1 - 0.64) \times 100 = 914$ (Eq2). In a third interaction between A and B, the expectation of individual A winning rises to $p = 0.73$ (Figure 1: Interaction 3, upper panel). If A wins again, this leads to $Elo_A = 1086 + (1 – 0.73) \times 100 = 1113$ and $Elo_B = 914 – (1 – 0.73) \times 100 = 887$ (Eq1 and Eq2). Note that the expected probability of A winning against B increases alongside the increasing difference between A's and B's ratings, while at the same time, the amount of points won and lost by each individual decreases (50, 36, 27, respectively). If however in a fourth interaction, B wins against A against the expectation (A is expected to win with $p = 0.79$), the amount of points gained and lost rises to 79, and the new ratings are $Elo_A = 1113 – 0.79 \times 100 = 1034$ (Eq4) and $Elo_B = 887 + 0.79 \times 100 = 966$ (Eq3, Figure 1: Interaction 4).

Appendix 2

The calculation of *S* is based on the assumption that it is justified to linearly extrapolate Elo ratings for days during which individuals were present but not observed. Therefore, *S* is clearly an approximate index.

We introduced a weighing factor to account for the notion that the higher in the hierarchy a rank change occurs, the more effect such a rank change has on stability. In other words, a rank reversal among the two highest individuals will have a stronger impact on the stability index than a rank reversal between the two lowest ranking individuals.

The weighing factor w_i, by which the sum of rank changes C_i is multiplied, is the standardized Elo rating of the highest rated individual involved in a rank change. Standardized Elo ratings are set between 0 and 1, for the lowest and highest rated individual present on a given day, respectively. Ratings of the remaining individuals are scaled in between. Thereby the differences between standardized and original ratings are proportional to each other. A rank reversal among the two highest individuals will therefore be weighed by $w_i = 1$, whereas a rank reversal among the two lowest individuals will be weighed by a value near 0. Please note that in the latter case the value of w_i depends on the standardized Elo rating of the second lowest rated individual and therefore does not equal 0.

Additionally, in case one individual leaves, we raised the ranks of all individuals below by one, thus defining $C_i = 0$ in such a case, given that rank changes other than those induced by one individual leaving the hierarchy did not occur.

References

Adamo, S. A. & Hoy, R. R. 1995. Agonistic behaviour in male and female field crickets, *Gryllus bimaculatus*, and how behavioural context influences its expression. *Animal Behaviour*, **49**, 1491-1501. doi:10.1016/0003-3472(95)90070-5.

Albers, P. C. H. & de Vries, H. 2001. Elo-rating as a tool in the sequential estimation of dominance strengths. *Animal Behaviour*, **61**, 489-495. doi:10.1006/anbe.2000.1571.

Altmann, J. 1974. Observational study of behavior: sampling methods. *Behaviour*, **49**, 227-267. doi:10.1163/156853974X00534.

Appleby, M. C. 1983. The probability of linearity in hierarchies. *Animal Behaviour*, **31**, 600-608. doi:10.1016/S0003-3472(83)80084-0.

Baayen, R. H. 2011. languageR: data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics", version 1.2, http://CRAN.R-project.org/package=languageR.

Bang, A., Deshpande, S., Sumana, A. & Gadagkar, R. 2010. Choosing an appropriate index to construct dominance hierarchies in animal societies: a comparison of three indices. *Animal Behaviour*, **79**, 631-636. doi:10.1016/j.anbehav.2009.12.009.

Bates, D. M., Maechler, M. & Bolker, B. M. 2011. lme4: Linear mixed-effects models using S4 classes, version 0.999375-39, http://CRAN.R-project.org/package=lme4.

Bayly, K. L., Evans, C. S. & Taylor, A. 2006. Measuring social structure: a comparison of eight dominance indices. *Behavioural Processes*, **73**, 1-12. doi:10.1016/j.beproc.2006.01.011.

Beehner, J. C., Bergman, T. J., Cheney, D. L., Seyfarth, R. M. & Whitten, P. L. 2005. The effect of new alpha males on female stress in free-ranging baboons. *Animal Behaviour*, **69**, 1211-1221. doi:10.1016/j.anbehav.2004.08.014.

Beletsky, L. D. 1992. Social stability and territory acquisition in birds. *Behaviour*, **123**, 290-313. doi:10.1163/156853992X00066.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.

Boyd, R. & Silk, J. B. 1983. A method for assigning cardinal dominance ranks. *Animal Behaviour*, **31**, 45-58. doi:10.1016/S0003-3472(83)80172-9.

Clutton-Brock, T. H., Albon, S. D., Gibson, R. M. & Guinness, F. E. 1979. The logical stag: adaptive aspects of fighting in red deer (*Cervus elaphus* L.). *Animal Behaviour*, **27**, 211-225. doi:10.1016/0003-3472(79)90141-6.

Côté, S. D. 2000. Dominance hierarchies in female mountain goats: stability, aggressiveness and determinants of rank. *Behaviour*, **137**, 1541-1566. doi:10.1163/156853900502718.

Côté, S. D. & Festa-Bianchet, M. 2001. Reproductive success in female mountain goats: the influence of age and social rank. *Animal Behaviour*, **62**, 173-181. doi:10.1006/anbe.2001.1719.

David, H. A. 1987. Ranking from unbalanced paired-comparison data. *Biometrika*, **74**, 432-436. doi:10.1093/biomet/74.2.432.

de Vries, H. 1995. An improved test of linearity in dominance hierarchies containing unknown or tied relationships. *Animal Behaviour*, **50**, 1375-1389. doi:10.1016/0003-3472(95)80053-0.

de Vries, H. 1998. Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. *Animal Behaviour*, **55**, 827-843. doi:10.1006/anbe.1997.0708.

de Vries, H. 2009. On using the DomWorld model to evaluate dominance ranking methods. *Behaviour*, **146**, 843-869. doi:10.1163/156853909X412241.

de Vries, H., Stevens, J. M. G. & Vervaecke, H. 2006. Measuring and testing the steepness of dominance hierarchies. *Animal Behaviour*, **71**, 585-592. doi:10.1016/j.anbehav.2005.05.015.

de Waal, F. B. M. & Luttrell, L. M. 1989. Toward a comparative socioecology of the genus *Macaca*: different dominance styles in rhesus and stumptail monkeys. *American Journal of Primatology*, **19**, 83-109. doi:10.1002/ajp.1350190203.

Dewsbury, D. A. 1990. Fathers and sons: genetic factors and social dominance in deer mice, *Peromyscus maniculatus*. *Animal Behaviour*, **39**, 284-289. doi:10.1016/S0003-3472(05)80872-3.

Drews, C. 1993. The concept and definition of dominance in animal behaviour. *Behaviour*, **125**, 283-313. doi:10.1163/156853993X00290.

Duboscq, J., Neumann, C., Perwitasari-Farajallah, D. & Engelhardt, A. 2008. Daytime birth of a baby crested black macaque (*Macaca nigra*) in the wild. *Behavioural Processes*, **79**, 81-84. doi:10.1016/j.beproc.2008.04.010.

Dubuc, C., Brent, L. J. N., Accamando, A. K., Gerald, M. S., MacLarnon, A. M., Semple, S., Heistermann, M. & Engelhardt, A. 2009. Sexual skin color contains information about the timing of the fertile phase in free-ranging *Macaca mulatta*. *International Journal of Primatology*, **30**, 777-789. doi:10.1007/s10764-009-9369-7.

East, M. L., Höner, O. P., Wachter, B., Wilhelm, K., Burke, T. & Hofer, H. 2009. Maternal effects on offspring social status in spotted hyenas. *Behavioral Ecology*, **20**, 478-483. doi:10.1093/beheco/arp020.

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

Engelhardt, A., Heistermann, M., Hodges, J. K., Nürnberg, P. & Niemitz, C. 2006. Determinants of male reproductive success in wild long-tailed macaques (*Macaca fascicularis*) - male monopolisation, female mate choice or post-copulatory mechanisms? *Behavioral Ecology and Sociobiology*, **59**, 740-752. doi:10.1007/s00265-005-0104-x.

Engh, A. L., Beehner, J. C., Bergman, T. J., Whitten, P. L., Hoffmeier, R. R., Seyfarth, R. M. & Cheney, D. L. 2006. Female hierarchy instability, male immigration and infanticide increase glucocorticoid levels in female chacma baboons. *Animal Behaviour*, **71**, 1227-1237. doi:10.1016/j.anbehav.2005.11.009.

Gammell, M. P., de Vries, H., Jennings, D. J., Carlin, C. M. & Hayden, T. J. 2003. David's score: a more appropriate dominance ranking method than Clutton-Brock et al.'s index. *Animal Behaviour*, **66**, 601-605. doi:10.1006/anbe.2003.2226.

Glickman, M. E. & Doan, T. 2010. The USCF Rating System, http://www.glicko.net/ratings/rating.system.pdf. accessed 2010-10-21.

Hamilton, W. J., III & Bulger, J. B. 1990. Natal male baboon rank rises and successful challenges to resident alpha males. *Behavioral Ecology and Sociobiology*, **26**, 357-362. doi:10.1007/BF00171102.

Holekamp, K. E. & Smale, L. 1991. Dominance acquisition during mammalian social development: the 'inheritance' of maternal rank. *American Zoologist*, **31**, 306-317. doi:10.1093/icb/31.2.306.

Hsu, Y., Earley, R. L. & Wolf, L. L. 2006. Modulation of aggressive behaviour by fighting experience: mechanisms and contest outcomes. *Biological Reviews*, **81**, 33-74. doi:10.1017/S146479310500686X.

Hvattum, L. M. & Arntzen, H. 2010. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, **26**, 460-470. doi:10.1016/j.ijforecast.2009.10.002.

Johnson, J. B. & Omland, K. S. 2004. Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101-108. doi:10.1016/j.tree.2003.10.013.

Kawamura, S. 1958. The matriarchal social order in the Minoo-B Group. *Primates*, **1**, 149-156. doi:10.1007/BF01813701.

Keiper, R. R. & Receveur, H. 1992. Social interactions of free-ranging Przewalski horses in semi-reserves in the Netherlands. *Applied Animal Behaviour Science*, **33**, 303-318. doi:10.1016/s0168-1591(05)80068-1.

Koenig, A. & Borries, C. 2006. The predictive power of socioecological models: a reconsideration of resource characteristics, agonism and dominance hierarchies. In: *Feeding*

Ecology of Apes and other Primates (Ed. by G. Hohmann, M. M. Robbins & C. Boesch), pp. 263-284. Cambridge: Cambridge University Press.

Kolmer, K. & Heinze, J. 2000. Rank orders and division of labour among unrelated cofounding ant queens. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 1729-1734. doi:10.1098/rspb.2000.1202.

Kurvers, R. H. J. M., Eijkelenkamp, B., van Oers, K., van Lith, B., van Wieren, S. E., Ydenberg, R. C. & Prins, H. H. T. 2009. Personality differences explain leadership in barnacle geese. *Animal Behaviour*, **78**, 447-453. doi:10.1016/j.anbehav.2009.06.002.

Landau, H. G. 1951. On dominance relations and the structure of animal societies: I. Effect of inherent characteristics. *Bulletin of Mathematical Biophysics*, **13**, 1-19. doi:10.1007/BF02478336.

Lange, H. & Leimar, O. 2004. Social stability and daily body mass gain in great tits. *Behavioral Ecology*, **15**, 549-554. doi:10.1093/beheco/arh044.

Lu, A., Koenig, A. & Borries, C. 2008. Formal submission, tolerance and socioecological models: a test with female Hanuman langurs. *Animal Behaviour*, **76**, 415-428. doi:10.1016/j.anbehav.2008.04.006.

Martin, P. & Bateson, P. 1993. *Measuring Behavior*, 2nd edn. Cambridge: Cambridge University Press.

Missakian, E. A. 1972. Genealogical and cross-genealogical dominance relations in a group of free-ranging rhesus monkeys (*Macaca mulatta*) on Cayo Santiago. *Primates*, **13**, 169-180. doi:10.1007/BF01840878.

Neumann, C., Assahad, G., Hammerschmidt, K., Perwitasari-Farajallah, D. & Engelhardt, A. 2010. Loud calls in male crested macaques, *Macaca nigra*: a signal of dominance in a tolerant species. *Animal Behaviour*, **79**, 187-193. doi:10.1016/j.anbehav.2009.10.026.

Pörschmann, U., Trillmich, F., Mueller, B. & Wolf, J. B. W. 2010. Male reproductive success and its behavioural correlates in a polygynous mammal, the Galapagos sea lion (*Zalophus wollebaeki*). *Molecular Ecology*, **19**, 2574-2586. doi:10.1111/j.1365-294X.2010.04665.x.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Rusu, A. S. & Krackow, S. 2004. Kin-preferential cooperation, dominance-dependent reproductive skew, and competition for mates in communally nesting female house mice. *Behavioral Ecology and Sociobiology*, **56**, 298-305. doi:10.1007/s00265-004-0787-4.

Rutte, C., Taborsky, M. & Brinkhof, M. W. G. 2006. What sets the odds of winning and losing? *Trends in Ecology & Evolution*, **21**, 16-21. doi:10.1016/j.tree.2005.10.014.

Sapolsky, R. M. 1983. Endocrine aspects of social instability in the olive baboon (*Papio anubis*). *American Journal of Primatology*, **5**, 365-379. doi:10.1002/ajp.1350050406.

Sapolsky, R. M. 2005. The influence of social hierarchy on primate health. *Science*, **308**, 648-652. doi:10.1126/science.1106477.

Schjelderup-Ebbe, T. 1922. Beiträge zur Sozialpsychologie des Haushuhns. *Zeitschrift für Psychologie*, **88**, 226-252.

Setchell, J. M., Smith, T. E., Wickings, E. J. & Knapp, L. A. 2008. Social correlates of testosterone and ornamentation in male mandrills. *Hormones and Behavior*, **54**, 365-372. doi:10.1016/j.yhbeh.2008.05.004.

Silk, J. B. 1993. Does participation in coalitions influence dominance relationships among male bonnet macaques? *Behaviour*, **126**, 171-189. doi:10.1163/156853993X00100.

Silk, J. B., Samuels, A. & Rodman, P. S. 1981. Hierarchical organization of female *Macaca radiata* in captivity. *Primates*, **22**, 84-95. doi:10.1007/BF02382559.

Smith, S. M. 1987. Responses of floaters to removal experiments on wintering chickadees. *Behavioral Ecology and Sociobiology*, **20**, 363-367. doi:10.1007/bf00300682.

Sprague, D. S., Suzuki, S., Takahashi, H. & Sato, S. 1998. Male life history in natural populations of Japanese macaques: migration, dominance rank, and troop participation of males in two habitats. *Primates*, **39**, 351-363. doi:10.1007/BF02573083.

Thierry, B. 2007. Unity in diversity: lessons from macaque societies. *Evolutionary Anthropology*, **16**, 224-238. doi:10.1002/evan.20147.

van Noordwijk, M. A. & van Schaik, C. P. 2001. Career moves: transfer and rank challenge decisions by male long-tailed macaques. *Behaviour*, **138**, 359-395. doi:10.1163/15685390152032505.

Vervaecke, H., Stevens, J. M. G., Vandemoortele, H., Sigurjónsdóttir, H. & de Vries, H. 2007. Aggression and dominance in matched groups of subadult Icelandic horses (*Equus caballus*). *Journal of Ethology*, **25**, 239-248. doi:10.1007/s10164-006-0019-7.

von Holst, D., Hutzelmeyer, H., Kaetzke, P., Khaschei, M., Rödel, H. G. & Schrutka, H. 2002. Social rank, fecundity and lifetime reproductive success in wild European rabbits (*Oryctolagus cuniculus*). *Behavioral Ecology and Sociobiology*, **51**, 245-254. doi:10.1007/s00265-001-0427-1.

Whitehead, H. 2008. *Analyzing Animal Societies: Quantitative Methods for Vertebrate Social Analysis*. Chicago: University of Chicago Press.

Widdig, A., Bercovitch, F. B., Streich, W. J., Sauermann, U., Nürnberg, P. & Krawczak, M. 2004. A longitudinal analysis of reproductive skew in male rhesus macaques. *Proceedings of the Royal Society B: Biological Sciences*, **271**, 819-826. doi:10.1098/rspb.2003.2666.

Wittemyer, G. & Getz, W. M. 2006. A likely ranking interpolation for resolving dominance orders in systems with unknown relationships. *Behaviour*, **143**, 909-930. doi:10.1163/156853906778017953.

Figure legends

Figure 1. Graphical illustration of Elo rating principles. Two individuals A (squares) and B (circles) interact four times out of which the first three interactions are won by A and the fourth is won by B. The amount of points gained/lost depends on the probability that the higher rated individual wins the interaction (see text for details). The winning probability (p) is a function of the difference in Elo ratings before the interaction (dotted vertical lines). As the difference in ratings increases with each interaction so does the chance of A winning. A graphical way to obtain the winning chance is depicted in the upper panel of the figure. A detailed description of this example can be found in appendix 1.

Figure 2. Elo ratings of ten male crested macaques during March 2007 (group R2). Each line represents one male. Each symbol represents Elo ratings after they were updated following an interaction of the depicted individual. Note that on March 10th, the residing top ranking male (SJ) and another high ranking male (YJ) emigrated from the group and a new male (ZJ) joined the group on March 11th, becoming the group's new alpha male (see text for details).

Figure 3. Elo ratings of eleven male crested macaques between June and August 2007 (group R2). Please note that the time scale differs from Figure 2 and for all males except KJ, symbols represent every 5th interaction (see text for details).

Figure 4. The development of dominance status of 16 natal male crested macaques during the six months before their emigration. Whereas using David's score only suggests an increase of status over time (a), Elo rating indicates a clear linear increase (b). Elo rating in addition allows a refinement of the time resolution, thereby suggesting a noticeable surge in ratings about three months before emigration (c, see text for details).

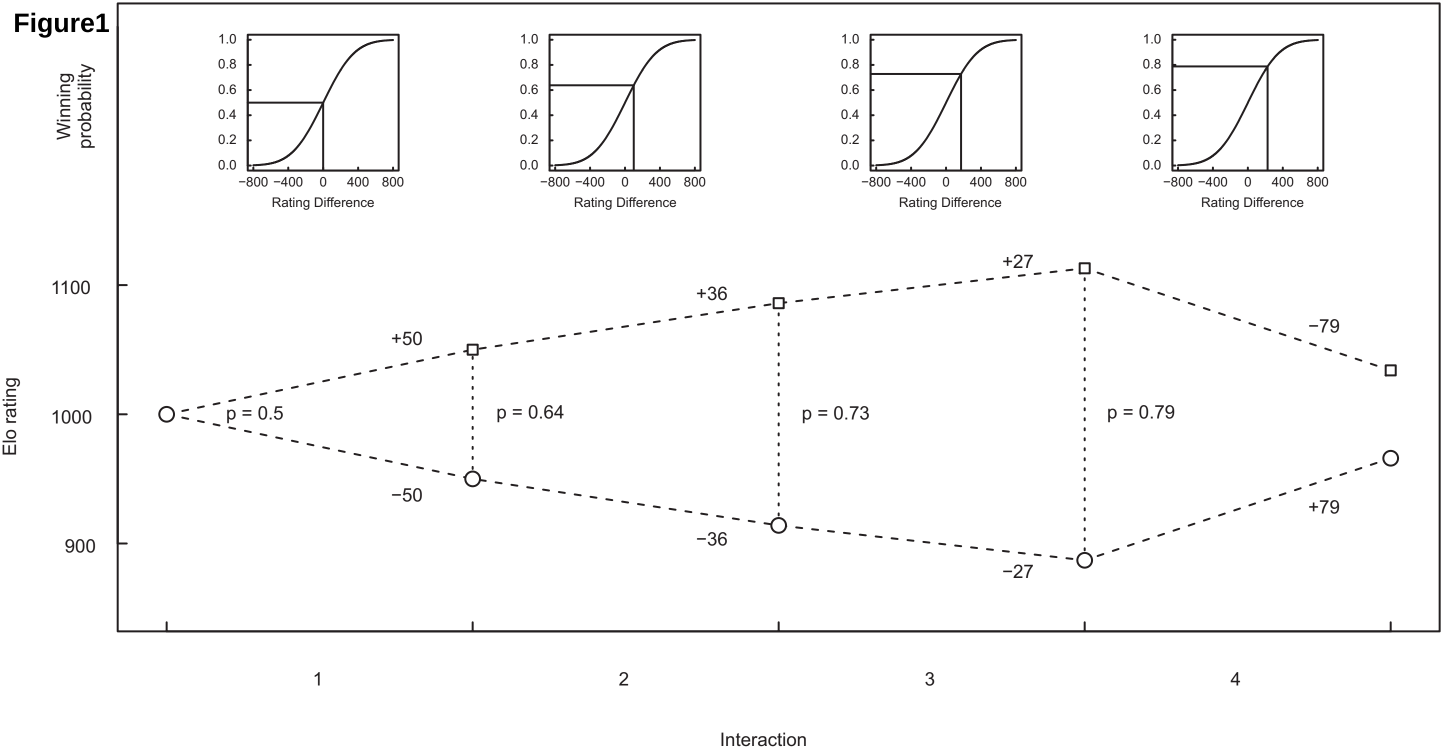
Figure 5. Correlation between the increase in unknown relationships and the performance of Elo rating, David's score and I&SI. The increase in unknown relationships was induced by randomly removing 50% of data points and performance is expressed as the correlation coefficient between rankings from the full and reduced data sets. Elo ratings and I&SI ranks are not influenced by higher percentages of unknown relationships, whereas the performance of David's score decreases when unknown relationships increase.

Acknowledgements

We thank Paul Albers, Frank W Albert, Han de Vries, Cedric Girard-Buttoz, Doreen Hess and Roger Mundry for discussions on dominance relationships and ranking methods, and James Higham for comments on the manuscript. Louise Barrett and two anonymous referees provided helpful criticism of the manuscript. We also thank all members of the Macaca Nigra Project that helped in collecting the behavioural data. We gratefully acknowledge the permission of the Indonesian State Ministry of Research and Technology (RISTEK), the Directorate General of Forest Protection and Nature Conservation (PHKA) and the Department for the Conservation of Natural Resources (BKSDA) in Manado, particularly Pak Untung and Pak Yunus, to conduct this research in the Tangkoko-Batuangus Nature Reserve. Furthermore, we thank the Caribbean Primate Research Center (CRPC) for permission to collect behavioral data on the rhesus macaque population at Cayo Santiago (Puerto Rico), and CPRC employees for their assistance during field work. The data collection on Cayo Santiago was approved by the IACUC of the University of Puerto Rico, Medical Sciences Campus (protocol No. 9400106 and 4060105, respectively). This project was funded by the DFG (grant No. WI 1808/3-1 to AW; and grant No. EN 719/2-1 to AE) and PhD stipends of the Saxon Ministry for Science and Culture (to CN), the Social Sciences and Humanities Research Council of Canada (to CD) and the Volkswagen Foundation (to JD). Cayo Santiago was supported by Grant Number CM-5 P40 RR003640-20 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

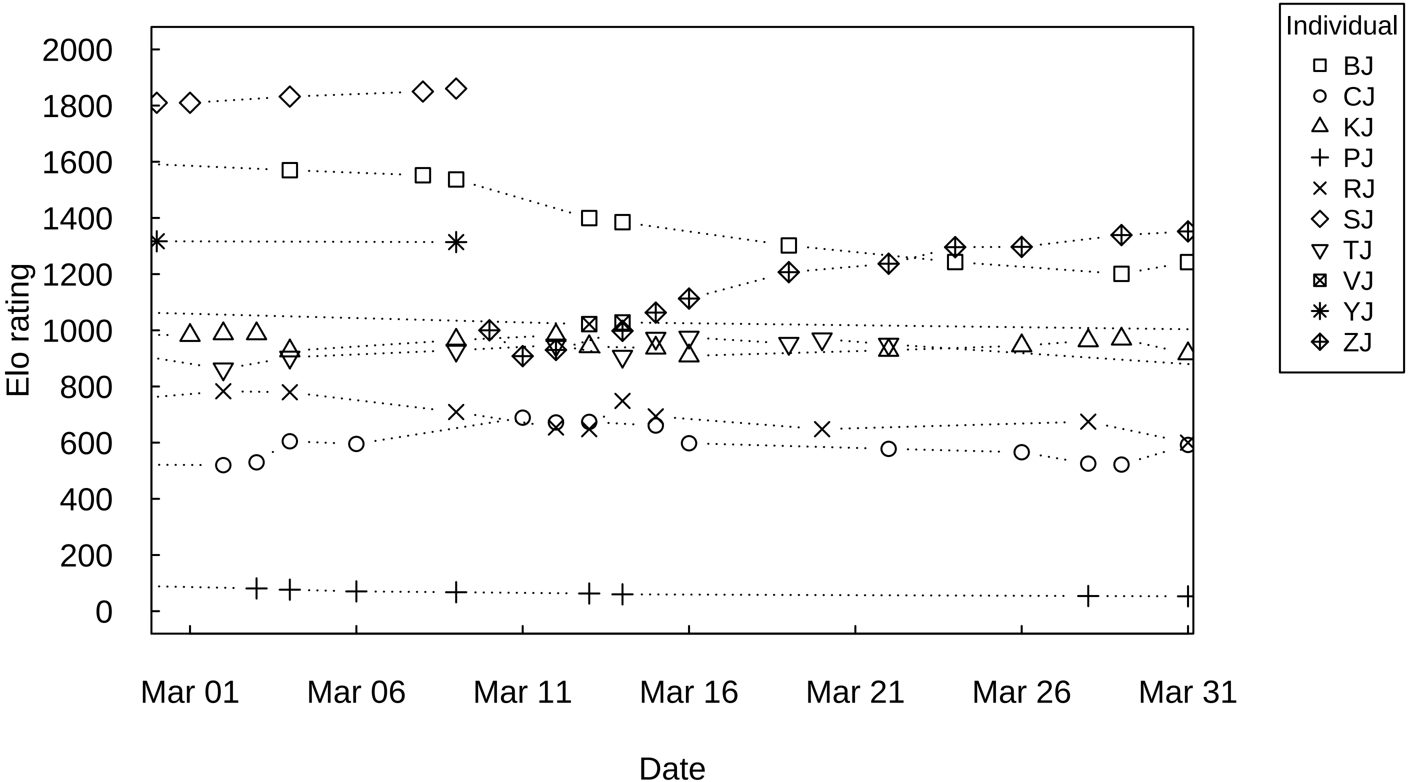
The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of NCRR or NIH.

Figure1



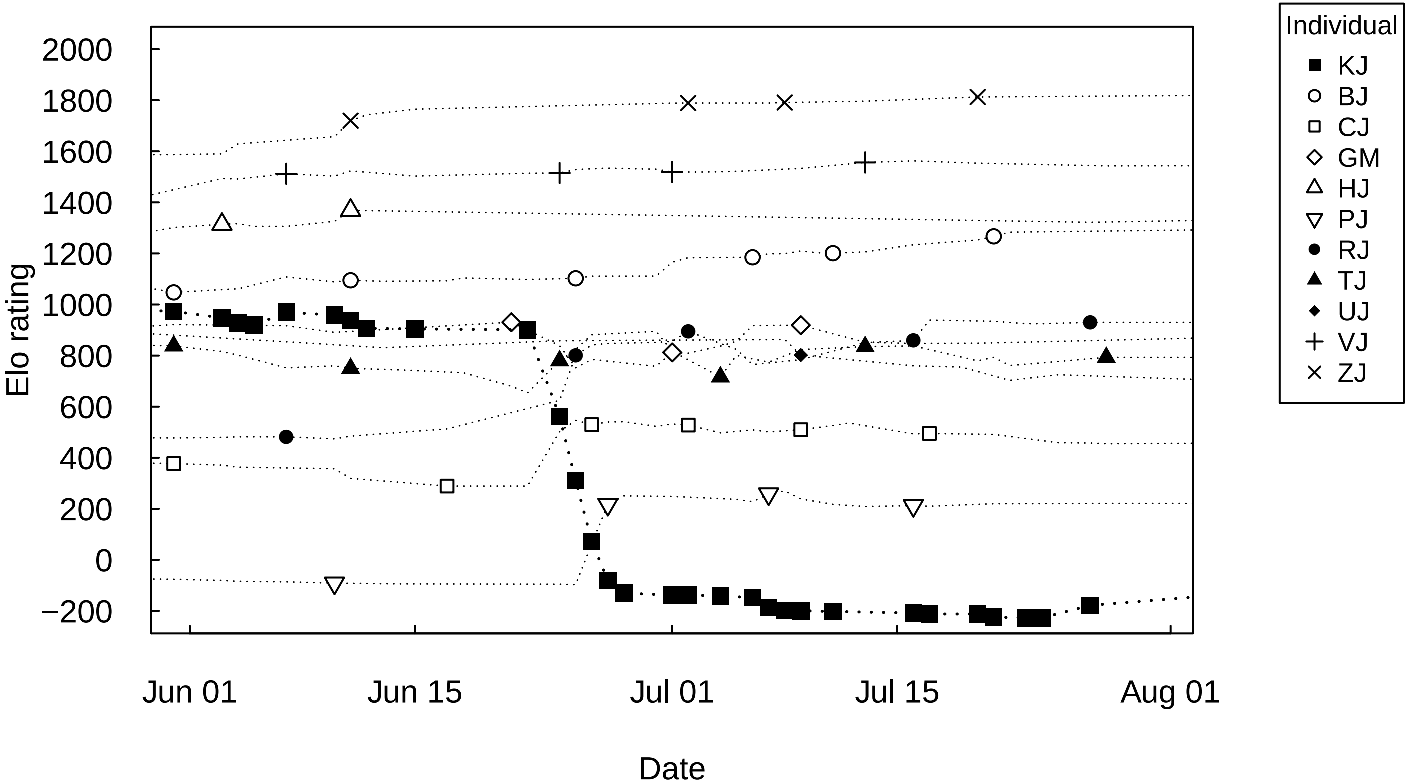
Winning probability
1.0
0.8
0.6
0.4
0.2
0.0
−800 −400 0 400 800
Rating Difference
Elo rating
1100
1000
900
+50
−50
p = 0.5
p = 0.64
+36
−36
p = 0.73
+27
−27
p = 0.79
−79
+79
1
2
3
4
Interaction

Figure2



Individual
BJ
CJ
KJ
PJ
RJ
SJ
TJ
VJ
YJ
ZJ
Elo rating
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Mar 01
Mar 06
Mar 11
Mar 16
Mar 21
Mar 26
Mar 31
Date

Figure3



Elo rating
2000
1800
1600
1400
1200
1000
800
600
400
200
0
−200
Jun 01
Jun 15
Jul 01
Jul 15
Aug 01
Date
Individual
KJ
BJ
CJ
GM
HJ
PJ
RJ
TJ
UJ
VJ
ZJ

Figure4

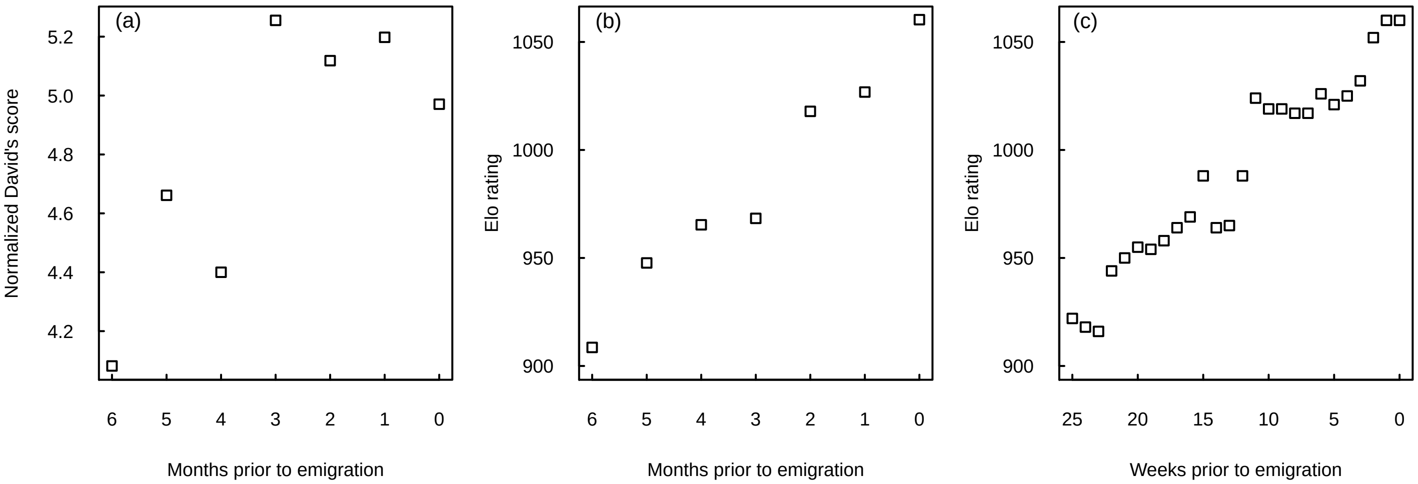
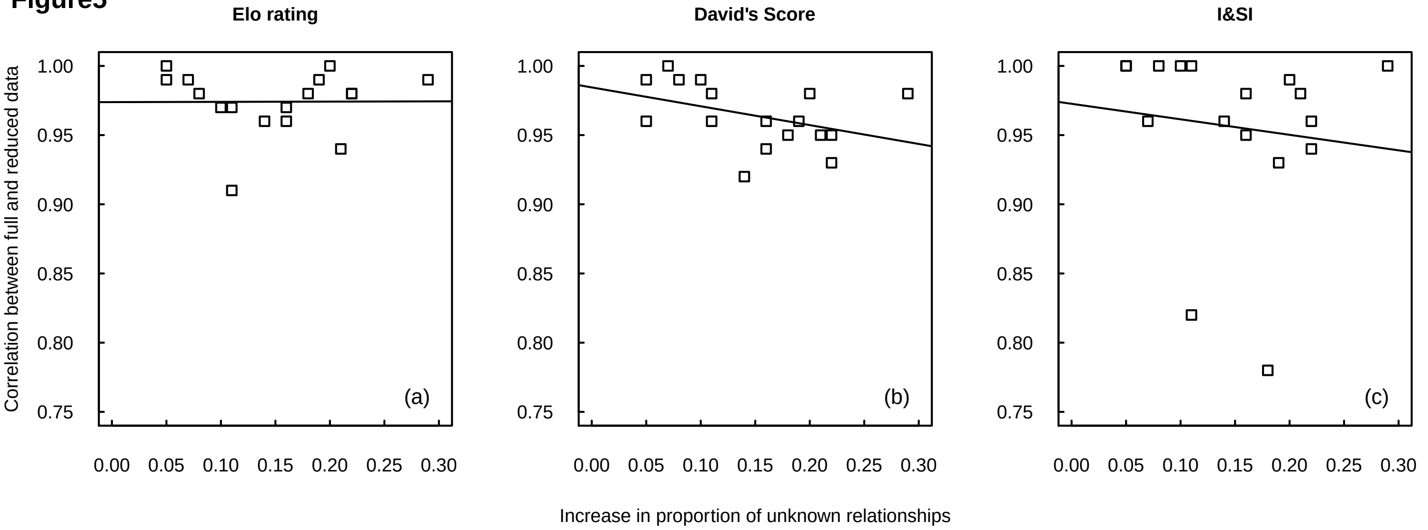

(a)
Normalized David's score
5.2
5.0
4.8
4.6
4.4
4.2
6
5
4
3
2
1
0
Months prior to emigration
(b)
Elo rating
1050
1000
950
900
6
5
4
3
2
1
0
Months prior to emigration
(c)
Elo rating
1050
1000
950
900
25
20
15
10
5
0
Weeks prior to emigration

Figure5



Elo rating
David's Score
I&SI
Correlation between full and reduced data
1.00
0.95
0.90
0.85
0.80
0.75
0.00
0.05
0.10
0.15
0.20
0.25
0.30
(a)
(b)
(c)
Increase in proportion of unknown relationships

Table 1. General description of the time periods and dominance matrices used in the analysis. Values are presented per species, group and sex. Average values are given as medians with inter-quartile ranges.

species	group	sex	*N* periods[a]	duration[b]	*N* individuals	Unknown relationships[c]		*N* interactions[d]
						proportion in full data set	increase in reduced data set	
mulatta	R	male	8	3.9 (3.1–4.1)	35 (34–42)	0.82 (0.79–0.88)	0.08 (0.06–0.09)	180 (123–234)
	V	female	4	1.8 (1.2–2.5)	22 (19–22)	0.66 (0.44–0.86)	0.13 (0.07–0.20)	116 (34–226)
		male	5	1.4 (1.1–2.9)	16 (16–20)	0.67 (0.58–0.71)	0.13 (0.12–0.14)	90 (41–125)
nigra	PB	female	3	4.0 (3.5–7.6)	18 (18–18)	0.25 (0.16–0.30)	0.19 (0.14–0.22)	299 (228–644)
		male	6	2.4 (2.2–	8 (7–9)	0.36 (0.25–	0.14 (0.11–	91 (50–112)

			3.5)		0.40)	0.16)	
R1	female	5	6.3 (5.8–11.2)	21 (21–22)	0.49 (0.47–0.57)	0.14 (0.07–0.16)	254 (158–292)
	male	16	2.6 (2.2–3.1)	10 (10–13)	0.34 (0.09–0.46)	0.16 (0.10–0.18)	159 (114–194)
R2	female	7	6.7 (4.8–7.5)	18 (16–20)	0.50 (0.45–0.56)	0.13 (0.11–0.15)	194 (136–246)
	male	12	3.1 (2.2–4.0)	8 (6–9)	0.26 (0.13–0.34)	0.10 (0.07–0.12)	64 (33–181)

[a] Number of time periods created

[b] Duration of time periods in months

[c] Proportion of unknown relationships in the full data matrices and the increase in proportion of unknown relationships in the reduced data set (see text)

[d] Number of agonistic interactions in each matrix

Table2

Table 1. Robustness analysis. Correlation coefficients (r_s) between rankings from full and reduced data sets. (Median and inter-quartile range)

Linearity[a]	*N*	Elo rating	David's score	I&SI
+	17	0.98 (0.97–0.99)	0.96 (0.95–0.98)	0.98 (0.95–1.00)
–	49	0.94 (0.89–0.98)	0.92 (0.86–0.95)	

[a] Linearity in the reduced data set: + linearity test yielded significant *h*' index, i.e., $P \leq 0.05$ (de Vries 1995); – linearity test did not yield significant *h*' index, i.e., $P > 0.05$