

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

3

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

6

7

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

16
17 The analysis of dominance has a long-standing history (Schjelderup-Ebbe 1922;
18 Landau 1951), and a great number of methods to assess hierarchies in animal societies
19 are currently available (reviewed in de Vries 1998; Bayly et al. 2006; Whitehead 2008).
20 Though differing in calculation complexity, all ranking methods presently used in studies
21 of behavioural ecology are based on interaction matrices. For this, a specific type of
22 behaviour or interaction, from which the dominance/subordination relationship of a given
23 dyad can be deduced, is tabulated across all individuals (see for example, Vervaecke et
24 al. 2007). This matrix can either be reorganized as a whole in order to optimize a
25 numerical criterion (e.g., I&SI: de Vries 1998; minimizing entries below the matrix
26 diagonal: Martin & Bateson 1993), or alternatively, an individual measure of success
27 calculated for each animal present (e.g., David's score: David 1987; CBI: Clutton-Brock
28 et al. 1979). In the latter case, a ranking can be generated by ordering the obtained
29 individual scores.

30

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

41

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

52

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

66

67 **Elo Rating Procedure**

68

69 Elo rating, in contrast to commonly used methods, is not based on an interaction
70 matrix, but on the sequence in which interactions occur. At the beginning of the rating
71 process, each individual starts with a predefined rating, for example a value of 1000. The
72 amount chosen here has no effect on the differences in ratings later: the relative distances
73 between individual ratings will remain identical (Albers & de Vries 2001). After each
74 interaction, the ratings of the two participants are updated according to the outcome of
75 the interaction: the winner gains points, the loser loses points. The amount of points

76 gained and lost during one interaction depends on the expectation of the outcome (i.e.,
77 the probability that the higher rated individual wins, Elo 1978) prior to this interaction.
78 Expected outcomes lead to smaller changes in ratings than unexpected outcomes (Figure
79 1). Depending on whether the higher rated individual wins or loses an interaction, ratings
80 are updated according to the following formulae:

81

82 Higher rated individual wins:

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

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

85

86 Lower rated individual wins (against the expectation):

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

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

89

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

99 even though the choice of k can have interesting implications (see section Integrity of
100 Power Assessment).

101

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

107

108 **Advantages of Elo Rating over Matrix Based Methods**

109 ***No minimum number of individuals***

110

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

120

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

123

124 ***Independence of Demographic Changes***

125

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

130

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

142

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

152

153 Since Elo rating does not stop the rating process as a consequence of changes in
154 group composition it circumvents a further drawback of matrix-based methods.
155 Techniques such as I&SI and David's score result in values that directly depend on the
156 number of individuals present, thus an observed change in calculated dominance rank or
157 score across two time periods may in fact be a consequence of changes in the number of
158 animals in the group rather than changes in competitive abilities, thus making a
159 comparison invalid. For example, in the case of the normalized David's score (c.f. de
160 Vries et al. 2006), values can range between 0 and $N - 1$, where N is the number of
161 individuals present in the social group. Elo rating, in contrast, results in ratings that do
162 not depend on the number of individuals present. Given that k is fixed for the entire rating
163 process, the current opponent's strength is the only variable that influences an
164 individual's future rating. Hence, the Elo rating of an individual is independent of the
165 number of individuals, and time periods that need to be created as a consequence of

166 changes in the number of individuals. This feature allows Elo rating to be used in a
167 longitudinal manner which is crucial for a wide array of studies, e.g., those on
168 mechanisms of rank acquisition and maintenance, determinants of life-time reproductive
169 success, and so on.

170

171 However, as in the other methods, true ratings of individuals are only known after
172 a minimum amount of interactions involving these individuals occurred (see also Albers
173 and de Vries 2001). For example (Figure 2), rank orders that would have been obtained
174 through Elo rating within the first two weeks of ZJ's group membership would have
175 placed him as ranking below BJ. After 13 days (i.e., eight observed interactions), ZJ
176 reached the top-ranked position in the Elo ratings. Using all observed interactions from
177 these two weeks it was not possible to construct a linear hierarchy, and only after 45 days
178 did we obtain a matrix with a sufficient amount of interactions permitting the use of
179 I&SI. However, it is likely that ZJ became alpha male directly upon his arrival in the
180 group even though he lost his very first observed interaction (top entry: see e.g., Sprague
181 et al. 1998) rather than constantly rising through the hierarchy. Albers and de Vries
182 (2001) suggest waiting for at least two interactions before assessing a dominance
183 hierarchy through Elo rating whenever a new member joins the hierarchy: one against a
184 stronger and one against a weaker opponent. In the case of ZJ, however, we observed him
185 interacting with six out of the seven other males present. In our case it thus seems more
186 appropriate to follow Glickman and Doan's (2010, rating chess players) suggestion to
187 treat ratings based on less than nine interactions as 'provisional' and exclude such ratings
188 from rankings. Therefore in general, Elo rating still needs a short start-up time before

189 creating reliable dominance hierarchies when group composition changes. This start-up
190 time is however much shorter than the time needed to build up sufficiently filled
191 interaction matrices for dominance hierarchies.

192

193 ***Visualization and Monitoring of Hierarchy Dynamics***

194

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

199

200 One of the great advantages of Elo rating is its ability to visualise dominance
201 relationships on a time scale, thus allowing monitoring of rank relationship dynamics. As
202 the information about the sequence of interactions is a prerequisite for applying Elo
203 rating, one can easily create graphs that depict the time scale on the x-axis and plot the
204 development of each individual's ratings on the y-axis. This approach can demonstrate a
205 fundamental feature of Elo rating, i.e., the possibility to obtain a rank order at any given
206 point in time by ordering the most recently updated ratings for a given set of individuals.
207 For example (Figure 2), the ordinal rank order among the present individuals on March
208 1st based on Elo ratings was SJ (1810 Elo points), BJ (1592), YJ (1317), VJ (1068), KJ
209 (982), TJ (942), RJ (703), CJ (526), PJ (90). By March 31st, however, the ordinal rank
210 order had changed into ZJ (1355), BJ (1262), VJ (994), TJ (950), KJ (892), RJ (600), CJ
211 (592), PJ (53).

212

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

220

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

232

233 In Old World monkeys and many other group living mammals, it is sometimes
234 observed that young males rise in rank before they eventually leave their natal group

235 (e.g., Hamilton & Bulger 1990). A common approach to quantify this phenomenon would
236 be to calculate monthly ranks and correlate them with the time to departure. Doing so for
237 16 natal male crested macaques (see below for details on the study population and data
238 collection) using David's score, however, lends only little support to this phenomenon
239 (Spearman's rank correlation: $r_s = 0.642$, $P = 0.139$, $N = 7$, Figure 4a). As described
240 below, this may be the consequence of high proportions of unknown relationships leading
241 to less reliable scores. It could also be due to the fact that David's scores directly depend
242 on the number of individuals incorporated in the matrix. In contrast, when using Elo
243 rating, the hypothesis that natal males rise in rank before emigration is strongly supported
244 ($r_s = 1$, $P < 0.001$, $N = 7$, Figure 4b). We observe an almost linear increase in ratings
245 before the migration date. It appears that males went through a noticeable surge about
246 three months before emigration, and kept rising before their departure. This is, however,
247 a preliminary result and further investigation is warranted. Since Elo ratings can be
248 obtained at any desired date, even an analysis with higher time resolution (e.g., weekly) is
249 possible (Figure 4c).

250

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

257

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

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

261 Eq5: ,
262 where C_i is the sum of absolute differences between rankings of two consecutive days, w_i
263 is a weighing factor determined as the standardized Elo rating of the highest ranked
264 individual involved in a rank change, and N_i is the number of individuals present on both
265 days (see appendix 2 for further details). Before division, values are summed over the
266 desired time period, i.e. n days. S can take values between 0, indicating a stable hierarchy
267 with identical rankings on each day of the analyzed time period, and $2 / \max(N_i)$,
268 indicating that the hierarchy is reversing every other day, i.e. total instability. Our data
269 suggest that S typically ranges between 0 and 0.5.

270

271 To test the validity of this approach we calculated S before and after the
272 immigration of male macaques that subsequently achieved high ranks (among the top
273 three, see below for details on the study population and data collection). We expected
274 such events to induce instability (e.g., Lange & Leimar 2004; Beehner et al. 2005), thus
275 leading to higher S values when compared to periods before such incidents. We found
276 less stability, i.e. greater S values, during four-week periods after the immigration of
277 males that achieved high rank compared to the four-week periods before (Wilcoxon
278 signed rank test: $V = 87$, $N = 14$, $P = 0.030$), indicating that hierarchies were less stable

279 after the immigration of a high ranking male. In contrast, after the immigration of males
280 that subsequently held low ranks, we observed no such difference in stability ($V = 14$, N
281 $= 7$, $P = 1.000$).

282

283 Such a quantitative approach may be advantageous since, so far, hierarchical
284 instability has been identified in a non-consistent manner. Sapolsky (1983) for example,
285 studying baboons, identified periods of instability in male dominance hierarchies through
286 high rates of ambiguously ending agonistic interactions and through high rates of
287 interactions that ended with the subordinate winning. In a different study of baboons,
288 Engh et al (2006) assessed instability in female dominance hierarchies in a mere
289 descriptive way. On a long-term basis, stability has also been characterised by
290 comparison of rankings in consecutive seasons using regression or correlation analysis
291 (e.g., in mountain goats, Côté 2000). By objectively defining stability, Elo rating may
292 become an important tool for studies on social instability and its consequences, for
293 example on individual stress levels and health (e.g., Sapolsky 2005), territory acquisition
294 (e.g., Beletsky 1992) or group transfer (e.g., Smith 1987; van Noordwijk & van Schaik
295 2001). In addition, the objective quantification of stability may make comparisons across
296 studies possible.

297

298 ***Independence of Time Periods***

299

300 It is common practice to obtain hierarchies at some arbitrary fixed time interval (e.g.
301 monthly). Given the dynamics of animal societies, both in group composition and

302 rankings (see above), such an approach is prone to misjudgement of hierarchies for two
303 reasons. First, all individuals incorporated in a dominance matrix must have the
304 possibility to interact with each other at all times. If group composition changes within
305 the studied interval, for example in fission/fusion societies or when individuals leave and
306 join frequently (floaters), applying matrix based methods is unjustified. Second, rank
307 changes that occur will be blurred (see the example above, Figure 3).

308

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

313

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

322

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

325

326 ***Integrity of Power Assessment***

327

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

333

334 *Undecided interactions*

335 Though some matrix-based methods (e.g., David's score or Boyd and Silk's
336 (1983) index) explicitly allow interactions without unambiguous winners and losers, i.e.,
337 draws or ties, to be taken into account when establishing dominance orders, researchers
338 (including us) usually choose to discard such observations. Clearly, agonistic interactions
339 that end without unambiguous winners and losers contain information about competitive
340 abilities of the involved individuals and should therefore not be disregarded. When using
341 Elo rating, an undecided interaction can be incorporated into the rating process to the
342 disadvantage of the higher rated individual whose rating will decrease, even though the
343 decrease will be smaller than had the higher rated individual lost the interaction (Albers
344 & de Vries 2001). After a draw the rating for the higher rated individual is reduced to
345 $\text{Rating}_{\text{new}} = \text{Rating}_{\text{old}} - k(p - 0.5)$, whereas the rating for the lower rated individual
346 increases to $\text{Rating}_{\text{new}} = \text{Rating}_{\text{old}} + k(p - 0.5)$. Hence, a draw between two individuals
347 that had identical ratings before the interaction (i.e., $p = 0.5$) will not alter the ratings. In

348 this way, Elo rating allows for a more complete power assessment of individuals by
349 including interactions into the rating process that are just as meaningful as clear winner-
350 loser interactions.

351

352 *Agonistic interactions of different quality*

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

359

360 *Choosing k*

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

370

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

376

377 **Testing the Reliability and Robustness of Elo Rating**

378

379 So far, we have shown how Elo-rating circumvents the problems associated with
380 matrix based methods. However, we have not yet shown how it compares to other
381 methods in terms of reliability and robustness. We now compare Elo-rating with two
382 widely used ranking methods that are based on interaction matrices (I&SI and David's
383 score), using our own empirical data. Mimicking a variety of social systems, we use data
384 collected on two species of macaques with different aggression patterns, crested (*Macaca*
385 *nigra*, aggressive interactions frequent, but of low intensity) and rhesus macaques (*M.*
386 *mulatta*, aggressive interactions less frequent, but of higher intensity) (de Waal & Luttrell
387 1989; Thierry 2007), and calculate dominance hierarchies for females (more stable
388 hierarchies) and males (more dynamic hierarchies) separately. To facilitate the
389 assessment of these analyses we will first briefly review the two methods we use for our
390 comparisons.

391

392 **Short Introduction to I&SI and David's Score**

393

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

403

404 David's score (David 1987) is an individual measure of success, in which for each
405 individual a score is calculated based on the outcome of its agonistic interactions with
406 other members of the social group as $DS = w + w_2 - l - l_2$, where w is the sum of an
407 individual's winning proportions and l the summed losing proportions. w_2 represents an
408 individual's summed winning proportions (i.e., w) weighed by the w values of its
409 interaction partners and likewise, l_2 equals an individual's summed losing proportions
410 (i.e., l) weighed by the l values of its interaction partners (David 1987; Gammell et al.
411 2003; see de Vries et al. 2006 for an illustrative example). Thus, David's score takes the
412 relative strength of opponents into account, valuing success against stronger individuals
413 more than success against weaker individuals.

414

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

417

418 **Methods**

419

420 *Study populations*

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

427

428 *Data collection*

429 Between 2006 and 2010, we collected data in three groups (R1, R2, PB) of a
430 population of wild crested macaques in the Tangkoko-Batuangus Nature Reserve, North
431 Sulawesi, Indonesia (1°33' N, 125 °10' E; e.g., Duboscq et al. 2008; Neumann et al.
432 2010). Groups comprised between 4 – 18 adult males and 16 – 24 adult females and were
433 completely habituated to human observers and individually recognizable. Between 2007
434 and 2010, data on rhesus macaques were collected in two groups (V, R) on the free
435 ranging population on Cayo Santiago, Puerto Rico (18°09' N, 65°44' W). The study
436 groups comprised between 20 – 60 females and 16 – 54 males (e.g., Dubuc et al. 2009,
437 Widdig unpublished data).

438

439 We collected data on dyadic dominance interactions, i.e., agonistic interactions
440 with unambiguous winner and loser, and displacement (approach / leave) interactions
441 during all occurrence sampling on focal animals and during ad libitum sampling
442 (Altmann 1974). Overall, our data set comprised a total of 12,740 interactions involving
443 252 individuals. Dominance hierarchies were created separately for the different species,
444 groups and sexes.

445

446 *Data analysis*

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

457

458 We tested all 66 matrices for linearity by means of de Vries' (1995) h' index. For
459 the 29 matrices for which the linearity test yielded a significant result, we applied de
460 Vries' (1998) I&SI method. Next, we calculated normalized David's scores from all

461 matrices following de Vries et al. 2006. Finally, we calculated Elo ratings from all
462 interactions in each of the group/sex combinations as a whole using Elo ratings on the
463 last day of each time period for the comparison with I&SI ranks and David's scores. Elo
464 ratings were calculated with 1000 as initial value and k was set to 100.

465

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

475

476 In a second analysis, we explored whether Elo rating is a robust method under
477 conditions of sparse data and whether the performance of Elo rating under such
478 conditions is systematically related to the percentage of unknown relationships in the
479 interaction matrix. Please note that a sparse matrix is not necessarily a matrix with a
480 higher proportion of unknown relationships. For example, a matrix in which each dyad
481 was observed five times and all entries are above the diagonal (i.e., there are no unknown
482 relationships) is more sparse than a matrix with each dyad being observed ten times
483 (likewise, no unknown relationships). Whereas the I&SI ranking will be identical in both

484 cases, David's scores will differ between the two, as will Elo ratings based on the
485 interactions leading to this matrix.

486

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

494

495 To explore the robustness of the method further, we tested whether Elo rating is
496 affected by increased proportions of unknown relationships and how it compared to the
497 two other methods. In other words, we investigated whether the methods become less
498 reliable as the proportion of unknown relationships increases. An increase in unknown
499 relationships was generated as a consequence of the random deletion of 50% of all
500 observed interactions (increase per period on average: 12.5%, inter-quartile range: 8 –
501 17%, "reduced data set": Table 1). We tested for an association between the increase in
502 unknown relationships and the correlation coefficient between ratings from the full and
503 reduced data set.

504

505 **Results**

506

507 Our results show that Elo ratings correlated highly with both I&SI ranks (median
508 $r_s = 0.97$, quartiles: 0.94–0.99, $N = 29$ periods) and David’s scores (median $r_s = 0.97$,
509 quartiles: 0.96–0.99, $N = 29$ periods).

510

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

517

518 Whereas there was no relationship between the increase in unknown relationships and
519 the correlation coefficient between full and reduced data sets for Elo rating ($r_s = -0.07$, N
520 $= 17$, $P = 0.799$) and I&SI ($r_s = -0.36$, $N = 17$, $P = 0.162$), we found that as the
521 proportion of unknown relationships increased the correlation coefficients decreased
522 between rankings from full and reduced data sets when using David’s score ($r_s = -0.52$, N
523 $= 17$, $P = 0.031$, Figure 5). Controlling for the initial proportion of unknown relationships
524 by means of a partial Spearman correlation test leads to similar results (Elo rating: $r_s = -$
525 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
526 $= 17$, $P = 0.006$),

527

528 Overall, our results indicate that Elo rating produces rank orders very similar to those
529 obtained with I&SI and David’s score. In addition, results of our tests suggest that

530 rankings from Elo rating and I&SI (given significant linearity test) remain stable in
531 sparse data sets, whereas David's score seems to create less reliable hierarchies in sparse
532 data sets as a result of an increase in unknown relationships.

533

534 ***Discussion***

535

536 Even though there is abundant literature available that compares the concordance of
537 different methods for the assessment of dominance ranks or scores (e.g., Bayly et al.
538 2006; Bang et al. 2010), this is the first study to test the reliability of Elo rating with an
539 extensive data set based on observations of free-ranging animals. Our results on
540 dominance interactions in crested and rhesus macaques show that Elo rating produces
541 dominance rank orders which closely resemble rankings generated with David's score
542 and the I&SI method. Furthermore, our results indicate that Elo rating is very robust
543 when data sets are limited in the number of interactions observed. Elo rating (and I&SI)
544 even seems to produce more reliable dominance hierarchies than David's score when the
545 proportion of unknown relationships is high. One could argue that this effect is due to the
546 initial proportion of unknown relationships, i.e., a relatively high proportion of unknown
547 relationships in a "full" matrix leads to some uncertainty in the ranking which may make
548 the scores from the further reduced matrix even less reliable. However, when controlling
549 for the initial proportion of unknown relationships, our results show that the robustness of
550 Elo rating (and I&SI) is not attributable to this factor.

551

552 Using Elo Rating – an Example

553

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

563

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

570

571 In addition to the three parameters that we originally found to be affected by
572 dominance rank, using Elo rating as predictor revealed two more acoustic parameters to
573 be significant at $P < 0.05$ (corrected for multiple testing after Benjamini and Hochberg
574 (1995), P values were assessed with the package languageR (Baayen 2011)). Using

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

581

582 **General Discussion**

583

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

591

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

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

603

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

618

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

622

623 **Appendix 1**

624

625 In this section, we give a detailed example of how Elo ratings are calculated.

626 Figure and equation references refer to the main article.

627

628 To illustrate the principles of Elo rating, it is useful to consider the basic unit of
629 any dominance hierarchy, the dyad. In the example presented here, two individuals A and
630 B interact through a sequence of four interactions. At the start of this sequence their
631 competitive abilities are unknown and thus there is no knowledge of their ratings, and
632 both A and B are assigned an initial rating of 1000. At this stage of the rating process,
633 both individuals are expected to be equally likely to win an interaction between each
634 other since there is not yet a higher rated individual, i.e., $p = 0.5$. If A wins the first
635 interaction against B, the ratings will be updated to $Elo_A = 1000 + (1 - 0.5) \times 100 = 1050$
636 (Eq1) and $Elo_B = 1000 - (1 - 0.5) \times 100 = 950$ (Eq2) (Figure 1: Interaction 1). Individual
637 A thus gained 50 points whereas B lost 50 points. Given that A has won the first
638 interaction, A is expected to win the next interaction against B with $p = 0.64$ due to the
639 rating difference between A and B of 100 (Figure 1: Interaction 2, upper panel). If A wins
640 the second interaction, ratings will be updated as follows: $Elo_A = 1050 + (1 - 0.64) \times 100$
641 $= 1086$ (Eq1) and $Elo_B = 950 - (1 - 0.64) \times 100 = 914$ (Eq2). In a third interaction

642 between A and B, the expectation of individual A winning rises to $p = 0.73$ (Figure 1:
643 Interaction 3, upper panel). If A wins again, this leads to $Elo_A = 1086 + (1 - 0.73) \times 100$
644 $= 1113$ and $Elo_B = 914 - (1 - 0.73) \times 100 = 887$ (Eq1 and Eq2). Note that the expected
645 probability of A winning against B increases alongside the increasing difference between
646 A's and B's ratings, while at the same time, the amount of points won and lost by each
647 individual decreases (50, 36, 27, respectively). If however in a fourth interaction, B wins
648 against A against the expectation (A is expected to win with $p = 0.79$), the amount of
649 points gained and lost rises to 79, and the new ratings are $Elo_A = 1113 - 0.79 \times 100 =$
650 1034 (Eq4) and $Elo_B = 887 + 0.79 \times 100 = 966$ (Eq3, Figure 1: Interaction 4).

651

652 **Appendix 2**

653

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

657

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

663

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

674

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

678

679 **References**

680

681 **Adamo, S. A. & Hoy, R. R.** 1995. Agonistic behaviour in male and female field crickets, *Gryllus*
682 *bimaculatus*, and how behavioural context influences its expression. *Animal Behaviour*, **49**, 1491-
683 1501. doi:10.1016/0003-3472(95)90070-5.

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

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

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

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

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

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

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

699 **Beehner, J. C., Bergman, T. J., Cheney, D. L., Seyfarth, R. M. & Whitten, P. L.** 2005. The
700 effect of new alpha males on female stress in free-ranging baboons. *Animal Behaviour*, **69**, 1211-
701 1221. doi:10.1016/j.anbehav.2004.08.014.

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

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

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

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

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

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

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

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

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

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

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

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

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

731 **Drews, C.** 1993. The concept and definition of dominance in animal behaviour. *Behaviour*, **125**,
732 283-313. doi:10.1163/156853993X00290.

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

736 **Dubuc, C., Brent, L. J. N., Accamando, A. K., Gerald, M. S., MacLarnon, A. M., Semple, S.,**
737 **Heistermann, M. & Engelhardt, A.** 2009. Sexual skin color contains information about the
738 timing of the fertile phase in free-ranging *Macaca mulatta*. *International Journal of Primatology*,
739 **30**, 777-789. doi:10.1007/s10764-009-9369-7.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

834 **Vervaecke, H., Stevens, J. M. G., Vandemoortele, H., Sigurjónsdóttir, H. & de Vries, H.**
835 2007. Aggression and dominance in matched groups of subadult Icelandic horses (*Equus*
836 *caballus*). *Journal of Ethology*, **25**, 239-248. doi:10.1007/s10164-006-0019-7.

837 **von Holst, D., Hutzelmeyer, H., Kaetzke, P., Khaschei, M., Rödel, H. G. & Schrutka, H.**
838 2002. Social rank, fecundity and lifetime reproductive success in wild European rabbits
839 (*Oryctolagus cuniculus*). *Behavioral Ecology and Sociobiology*, **51**, 245-254.
840 doi:10.1007/s00265-001-0427-1.

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

843 **Widdig, A., Bercovitch, F. B., Streich, W. J., Sauermaun, U., Nürnberg, P. & Krawczak, M.**
844 2004. A longitudinal analysis of reproductive skew in male rhesus macaques. *Proceedings of the*
845 *Royal Society B: Biological Sciences*, **271**, 819-826. doi:10.1098/rspb.2003.2666.

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

849

850

851 Figure legends

852

853 Figure 1. Graphical illustration of Elo rating principles. Two individuals A (squares) and
854 B (circles) interact four times out of which the first three interactions are won by A and
855 the fourth is won by B. The amount of points gained/lost depends on the probability that
856 the higher rated individual wins the interaction (see text for details). The winning
857 probability (p) is a function of the difference in Elo ratings before the interaction (dotted
858 vertical lines). As the difference in ratings increases with each interaction so does the
859 chance of A winning. A graphical way to obtain the winning chance is depicted in the
860 upper panel of the figure. A detailed description of this example can be found in appendix
861 1.

862

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

869

870

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

874

875

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

881

882

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

890

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.

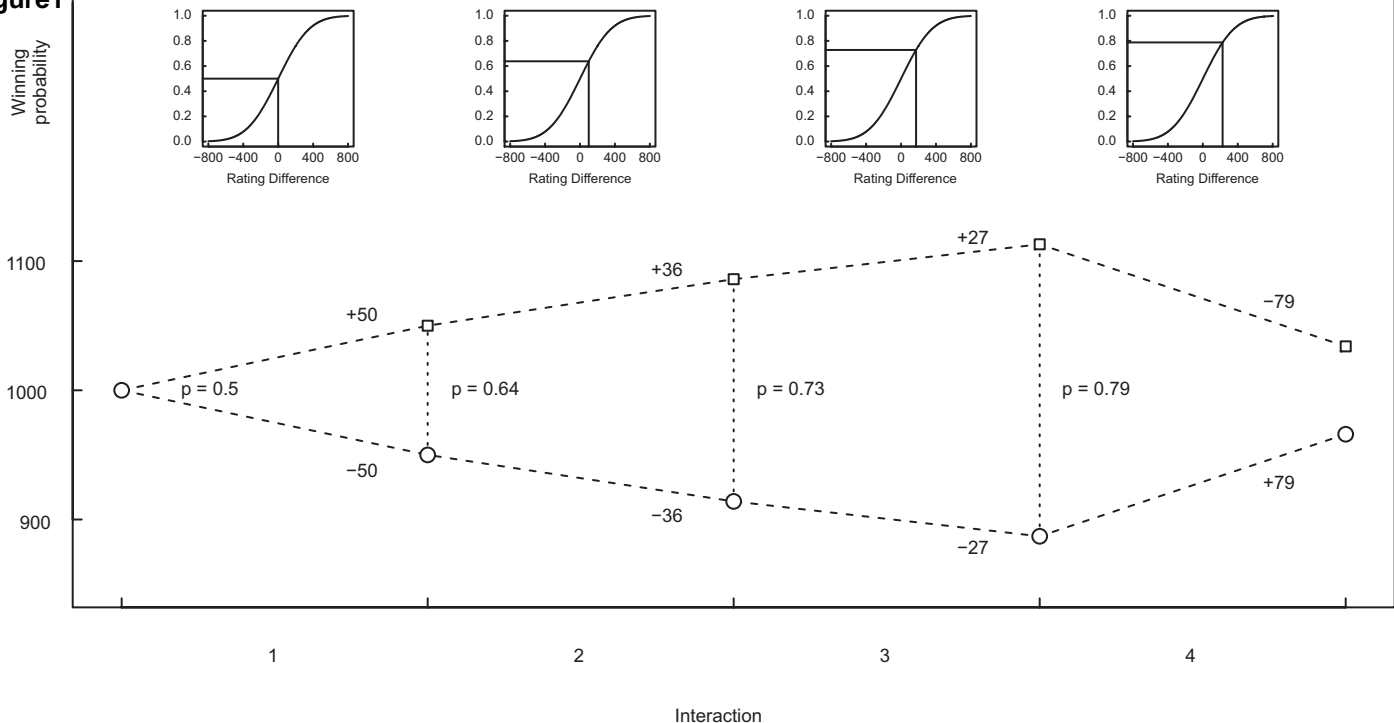
Figure 1

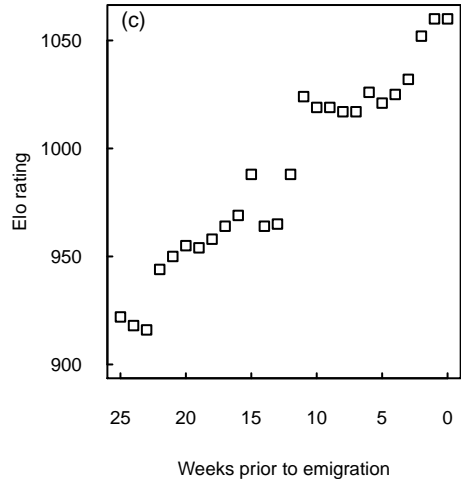
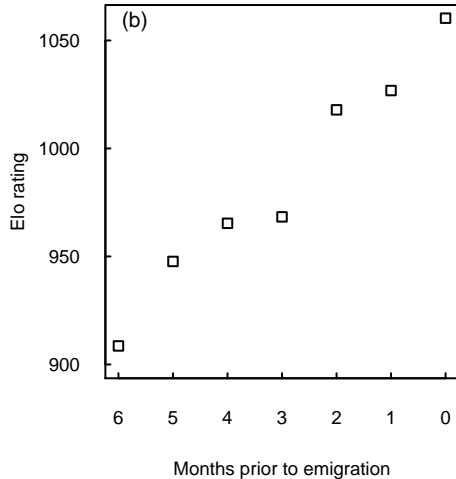
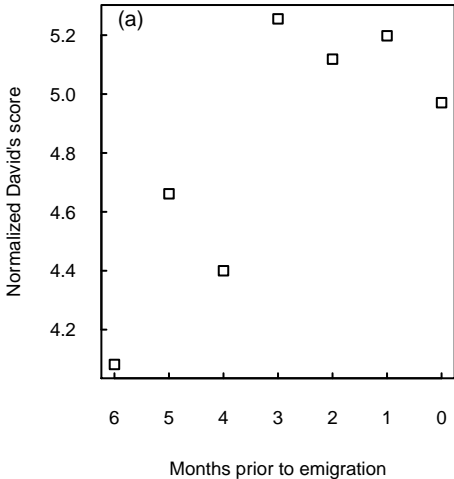
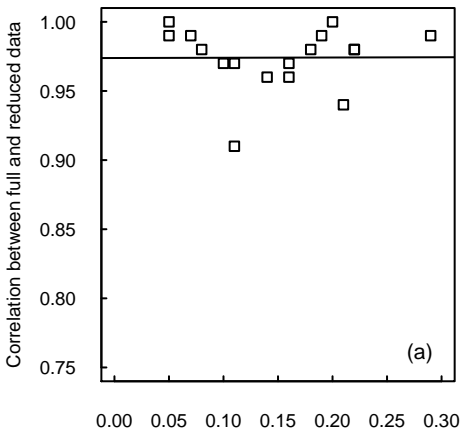
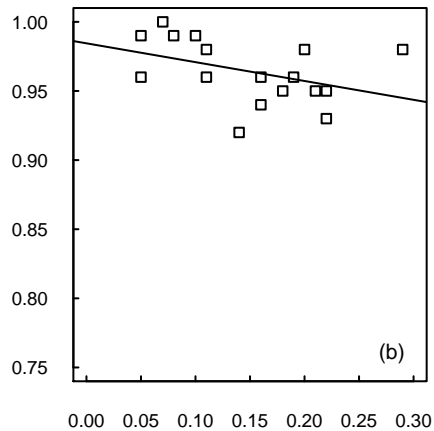
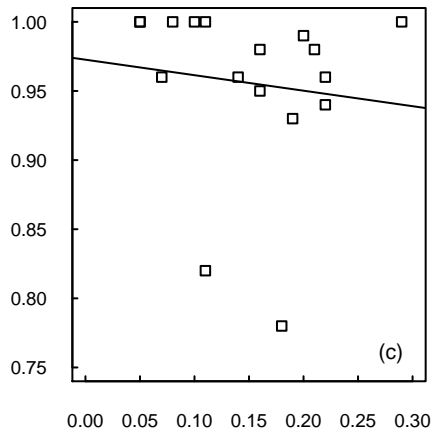
Figure4

Figure5**Elo rating****David's Score****I&SI**

Increase in proportion of unknown relationships

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

4

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

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

5 ^a Number of time periods created

6 ^b Duration of time periods in months

7 ^c Proportion of unknown relationships in the full data matrices and the increase in
8 proportion of unknown relationships in the reduced data set (see text)

9 ^d Number of agonistic interactions in each matrix

10

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

Linearity ^a	<i>N</i>	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., $P > 0.05$