Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies

Christof Neumann¹,² | Julia Fischer¹,²,³

¹Cognitive Ethology Laboratory, German Primate Center, Göttingen, Germany
²Leibniz Science Campus Primate Cognition, Göttingen, Germany
³Department of Primate Cognition, Georg-August-Universität Göttingen, Göttingen, Germany

Correspondence
Christof Neumann, Cognitive Ethology Laboratory, German Primate Center, 37077 Göttingen, Germany
Email: cneumann@dpz.eu; christofneumann1@gmail.com

Funding information
Deutsche Forschungsgemeinschaft,
Grant/Award Number: 254142454

Handling Editor: Matthew Silk

Abstract
1. The steepness of dominance hierarchies provides information about the degree of competition within animal social groups and is thus an important concept in socioecology. The currently most widely used metrics to quantify steepness are based on David’s scores (DS) derived from dominance interaction networks. One serious drawback of these DS-based metrics is that they are biased, that is, network density systematically decreases steepness values.
2. We provide a novel approach to estimate steepness based on Elo-ratings, implemented in a Bayesian framework (STEER: Steepness estimation with Elo-rating). We evaluate and validate its performance by means of experimentation on empirical and artificial datasets and compare its performance to that of several other steepness estimators.
3. STEER has two key advantages. First, it is unbiased, precise and more robust to data density than DS-based steepness. Second, it provides explicit probability distributions of the estimated steepness coefficient, which allows uncertainty assessment. In addition, it relies on the same underlying concept and is on the same scale as the original measure, and thus allows comparison to existing published results.
4. STEER provides a considerable improvement over existing methods to estimate dominance hierarchy steepness. We demonstrate its application with an example comparing within- and between species variation in steepness in a comparative analysis and present guidelines on how to use it. The R package EloSteepness allows convenient numeric and graphical assessment of the new steepness measure.

KEYWORDS
Bayesian methods, competition, dominance hierarchy, Elo-rating, social network
1 | INTRODUCTION

Analysing dominance relationships, dominance ranks and dominance hierarchies is a staple in studies of animal behaviour. Results of such analyses feature prominently in the description of social structure in many animal species, spanning vertebrates as well as invertebrates.

One key aspect in this context is hierarchy steepness, which can be defined as ‘the size of the absolute differences between adjacently ranked individuals in their overall success in winning dominance encounters’ (de Vries et al., 2006, p. 585). A hierarchy is considered steep if these differences are large, and a shallow hierarchy is one in which these differences between individuals are small. Hierarchy steepness is therefore often also referred to as dominance gradient (e.g. Barrett et al., 1999), and systems with steep hierarchies are often termed despotic while shallow systems are referred to as egalitarian, or less despotic (Sterck et al., 1997; Vehrencamp, 1983).

Steepness is particularly relevant for questions related to socioecology, dominance styles, biological markets and phylogenetic covariation of social traits (Amici et al., 2020; Balasubramaniam et al., 2012; Flack & de Waal, 2004; Schino & Aureli, 2008; Sterck et al., 1997; van Schaik, 1989).

For example, Balasubramaniam et al. (2012) investigated variation in hierarchy steepness in groups of nine species of macaque (genus Macaca). They found evidence for a strong phylogenetic signal in hierarchy steepness, which is exceptional given the generally low magnitude of phylogenetic signals in behavioural traits (Blomberg et al., 2003; Kamilar & Cooper, 2013). Their results also fit a wider literature on phylogenetic covariation in a suite of behavioural traits in this genus (Thierry et al., 2000, 2008). This result helped to reconcile the influence of both, phylogenetic history and environmental factors, in shaping variation in behavioural patterns across species (Balasubramaniam et al., 2012).

Quantifying hierarchy steepness has predominantly been done with one index (referred to here as ‘classic steepness’, de Vries et al., 2006). This index, however, suffers from one important drawback. Specifically, the steepness index decreases as network density decreases, that is, the more dyads that have no observed interactions in the dominance network, the shallower the hierarchy steepness (e.g. figure 2 in Klass & Cords, 2011, see Supplement S5, section 1). In other words, the higher the proportion of unknown relationships in a dataset (the sparser the network is), the smaller the steepness index becomes (see also Balasubramaniam et al., 2012).

A new or improved index that does not suffer from this issue is therefore desirable and should have at least the following properties. First, it should indeed capture the phenomenon it is supposed to capture, that is, the average difference in power differentials among individuals. Second, it should be robust to network density. An additional desirable feature would be that the index also assesses and incorporates the uncertainty arising from varying data density (i.e. the number of observed interactions per dyad regardless of network density, see Supplement S5, section 1). Here we propose a novel index to quantify hierarchy steepness that meets these three criteria and present an accompanying R package to handle its calculation and visualization. We begin by briefly describing the original steepness index of de Vries et al. (2006), followed by a description of our proposed novel index.

1.1 | Classic steepness based on David’s scores

Formally, steepness has been quantified as the slope of a regression of cardinal dominance scores of individuals on their ordinal dominance ranks, where the cardinal dominance scores are typically normalized David’s scores (de Vries et al., 2006). Calculating classic steepness (per de Vries et al., 2006) starts from a square matrix in which dyadic dominance interactions are tabulated. These raw interaction frequencies are then transformed into dyadic win/loss proportions \( P_{ij} \). For example, if the dyad AB interacted 10 times, and A won 9 of these interactions and lost 1, the winning proportion of A is \( P_{w_{A}} = \frac{9}{10} = 0.9 \), and its losing proportion is \( P_{l_{A}} = \frac{1}{10} = 0.1 \). Proportions for individual B calculated analogously. These winning and losing proportions are then summed for each individual. In the second step, these two sums are summed again while being weighted by the corresponding winning and losing proportions of each individual’s opponents. The final score for an individual then is the sum of its two winning values minus its two losing values (see de Vries et al. (2006) for more details). The result of this procedure is a single score for each individual (David’s score, David, 1987; Gammell et al., 2003) where high values indicate high success (‘high rank’) and low scores indicate low success (‘low rank’). These David’s scores are then normalized such that the scores of all individuals range between 0 and \( n-1 \), where \( n \) is group size. To derive the steepness metric, a simple linear regression is fit between scores and ordinal ranks of the scores (figure 1 in de Vries et al. (2006), see also Figure 3). The absolute value of the slope coefficient of this model then is the steepness metric. Because of the normalization step, steepness ranges between 0 and 1.

A variant of this approach also suggested by de Vries et al. (2006) uses dyadic proportions that are corrected for chance \( D_{ij} \), rather than the pure dyadic winning proportions \( P_{ij} \), to derive David’s scores. Throughout the evaluation of our new method, we consider both variants of classic steepness.

1.2 | Elo-rating-based steepness

Our approach to obtain a steepness index follows the same general logic: We use dyadic interactions to derive normalized individual dominance scores, which are used to fit a simple regression model, the slope of which then is the steepness metric. The key new features are that the individual scores are derived from Elo-ratings rather than being based on David’s scores and that we model steepness in a Bayesian framework, which results in posterior distributions rather than point estimates of steepness.

In brief, Elo-rating in its original form updates individual ratings after single observed dyadic interactions, where interaction winners
increase their ratings and losers decrease their ratings (Albers & de Vries, 2001; Elo, 1978; Goffe et al., 2018; Neumann et al., 2011). The magnitude of change in the ratings depends on the expected outcome of the interaction, which, in turn, depends on the rating difference between the two interactants prior to the interaction. For example, if individuals A and B interact and prior to their interaction, A had a much larger rating than B, the expectation is that A is very likely to win the interaction and conversely B is very unlikely to win. If A indeed wins the interaction, then the updated ratings will change very little for A and B. If, contrary to the expectation, B wins the interaction, ratings will change substantially.

In addition to the expected outcome, the exact amount by which ratings change depends on the parameter \( k \), which determines the maximum amount of change in ratings after a single interaction (Franz et al., 2015; Goffe et al., 2018). Typically, \( k \) is unknown and set to some arbitrary value like 100 (e.g. Neumann et al., 2011; see also below).

Importantly, the Elo-rating algorithm treats interactions sequentially, that is, individual ratings are updated after each interaction in the temporal order in which they occur/are observed. This constitutes a major difference to static methods like David’s score where the sequence of interactions is ignored. However, since we want to compare the performance of our Elo-rating-based method to that of static methods, we employ a randomization approach. For this, we translate static interaction matrices into randomized sequences in which the interactions may have occurred. This is necessary because the actual sequence of interactions is not available in static networks, which is the case when matrices are the data source [see also Sánchez-Tójar et al. (2018) and Clark et al. (2018)]. Note that our approach also works with data where the sequence of interactions is known. We did however focus our descriptions and evaluations on the case where there is no information about the sequence because our benchmark method (classic steepness) is implemented in such a framework.

For our purposes, we take advantage of the fact that with Elo-rating, we can express the expected winning probability for any individual with any potential opponent at any point in the rating process, which is simply a function of the differences in ratings between the two individuals. Importantly, these expected winning probabilities are defined regardless of whether two individuals actually interacted, which is yet another crucial difference to other ranking methods like David’s score.

It is also important to note that there is no consensus about the exact shape of the relationship between rating difference and winning probability (e.g. Franz et al., 2015; Goffe et al., 2018; Neumann et al., 2011; Sánchez-Tójar et al., 2018), although all implementations have in common a sigmoidal shape. Here we follow Goffe et al. (2018) and define the expected winning probability of A against B as

\[
p_{AB} = \frac{1}{1 + \exp(-k(\text{rating}_A - \text{rating}_B))}
\]

For example, consider the four individuals in Figure 1. The ratings in the top row are the ratings after all interactions in the sequence have been evaluated. Here A has the highest rating among the four individuals and we can calculate A’s expected winning probabilities with the three remaining individuals. Since A’s rating is the highest rating of all individuals, all its winning probabilities are larger than 0.5 (dashed red arrows in Figure 1), and in fact turn out to be all larger than 0.8.

For example, for individuals A and C, with ratings of 1.82 and –0.81, respectively, we find that

\[
p_{AC} = \frac{1}{1 + \exp(-0.81 - 1.82)} = 0.93
\]

FIGURE 1 From individual ratings to summed winning probabilities. From individual ratings (top panel), rating differences are calculated. These differences translate (right plot) into dyadic winning probabilities, which can be tabulated (left matrix). Winning probabilities are then summed for each individual to derive summed winning probabilities (bottom panel). Two individuals are highlighted (A in dashed red, C in full gold). The winning probabilities of each individual against itself are omitted (see main text).
Individual C, on the other hand, is expected to win only against one other individual (D) and is expected to lose against A and B. This translates into one expected winning probability larger than 0.5 and two smaller than 0.5 (full golden arrows in Figure 1).

These expected winning probabilities are then tabulated (centre left panel in Figure 1) per individual and opponent. When adding up the winning probabilities per individual we obtain **summed winning probabilities**. These summed winning probabilities are akin to normalized David’s scores in that they range between 0 for an individual that has winning probabilities of 0 against all other individuals and one for an individual that is expected to win against all other individuals, where n is group size. More formally, we define individual i’s summed winning probability $s_i$ as

$$s_i = \left( \sum_{j=1}^{n} \frac{1}{1 + \exp(r_j - r_i)} \right) - 0.5,$$

where $r_i$ is individual i’s rating at the end of the interaction sequence, that is, after all interactions were evaluated, $r_j$ is individual j’s rating at the end of the interaction sequence and n is the group size. We need to subtract 0.5 from this sum to account for the winning probability when $i = j$, that is, the winning probability of individual i against itself, which is $1 / (1 + \exp(0)) = 0.5$ and irrelevant.

With these summed winning probabilities at hand, we fit a regression model analogously to de Vries et al. (2006) with summed winning probabilities as a function of ordinal ranks of summed winning probabilities. The absolute value of the regression slope then is the steepness index (see also Figure 3).

### 1.3 | Tackling uncertainty

So far, we defined Elo-based steepness as a point estimate. There are however several sources of uncertainty for obtaining this estimate. First, we usually do not know the ratings of all individuals at the start of the interaction sequence. Second, we do not know the value of $k$, which determines how much exactly ratings change after each interaction. Both these issues have been addressed before (e.g. Franz et al., 2015; Goffe et al., 2018; Newton-Fisher, 2017), and we follow Goffe et al. (2018), who adopted an explicitly Bayesian approach to estimate start ratings and k from the available interaction data, rather than setting them to arbitrary values. Consequently, all quantities of interest, such as start ratings and k can be seen as probability distributions rather than fixed point estimates. The same is true for the actual individual ratings, expected dyadic winning probabilities and, in extension, to the summed winning probabilities of individuals and ultimately our new steepness measure.

A third source of uncertainty concerns the actual sequence in which the interactions occurred. Recall that we translate static matrices, that is, interactions that were aggregated over some time frame, into ‘dynamic’ interaction sequences to apply Elo-rating (Clark et al., 2018; Sánchez-Tójar et al., 2018). As we do not know the actual sequence, we simply use randomized versions of sequences in which the interactions could have occurred. The resulting multiple probability distributions can then simply be combined in the same way we combine posteriors from different chains during Markov chain Monte Carlo (MCMC) sampling.

A final source of uncertainty is data density. We refer to the mean number of observed interactions per dyad (in network terms: mean weight) as data density. More observations should lead to narrower estimates of individual scores and downstream to narrower estimates of steepness. We distinguish network density (proportion of unknown relationships or sparseness) from data density. Two interaction networks can have the same network density and group size, but differ in data density (see Supplement S5, section 1).

The concept is illustrated in Figure 2. Beginning from an interaction matrix, we obtain summed winning probabilities. In contrast to Figure 1, these are now distributions rather than point estimates, stemming from MCMC samples. The final step in obtaining steepness is then to rank the summed winning probabilities and fit a simple linear regression. This procedure is applied in each of the MCMC samples separately and results in a probability distribution of steepness values.

Figure 2 also shows that the amount of data (i.e. data density) directly informs how uncertain we need to be regarding the steepness estimates. Distributions shown in the bottom row vs. the top row of Figure 2 are based on the exact same win/loss ratios, but differ with respect to the number of observed interactions: the bottom row has twice the number of observed interactions. With more observed interactions, the resulting distributions become narrower, as they should. The more information (here: observed interactions per dyad) we have, the less uncertain we can be about our estimates. Another way of looking at this is that the summed winning probability distributions in the lower panel overlap much less compared to the upper panel.

It is also noteworthy that using the Bayesian Elo-model in this context handles ambiguous cases very naturally. In the example here, individuals d and e have a tied relationship, that is, they both won and lost the same number of interactions with each other. As a result, their summed winning probability distributions overlap to a very large extent.$^2$

In summary, our new steepness measure follows an analogous approach as classic steepness by using standardized scores of individuals as the source to estimate the steepness slope (Figure 3). In contrast to classic steepness though, these individual scores represent probability distributions of summed winning probabilities derived from Elo-ratings. We therefore refer to it as STEER: steepness estimation with Elo-rating. STEER captures uncertainty on multiple levels: first, the uncertainty arising from the sequence itself, that is, by randomizing the order in which interactions are considered; second, from the actual rating process, that is, using Bayesian estimates of k and start ratings (Goffe et al., 2018).

### 2 | MATERIALS AND METHODS

We used two datasets, which we describe in more detail below. The first was a set of artificially generated matrices. The second was a set of empirical datasets, that is, matrices extracted from published sources.
With these datasets, we used two complementary approaches to evaluate the performance of STEER in comparison to the original (classic) steepness measure and three other options to quantify steepness.

First, we looked at how well the different methods recovered some underlying (true) steepness value. This analysis is concerned only with the artificial data, because here we knew ground truth as we set the steepness parameter during data generation (see below and Supplement S5, section 2). Given the known detrimental effects of unknown relationships on this steepness measure (Klass & Cords, 2011), we only used dense interaction networks (less than...
5% unknown relationships) for this analysis. An ideal method would have had a correlation $\rho = 1$ (‘precise’) and a regression slope $\beta = 1$ (‘unbiased’) when looking at the relationship between ground truth and the results of the different methods to assess steepness.

Second, we investigated how well the different methods dealt with varying network density. Here, we performed a removal experiment, in which we increased the proportion of unknown relationships incrementally in a given dataset by removing interactions, and subsequently quantified the relationship between steepness and the proportion of unknown relationships in each dataset. This approach provided insights into the robustness of the different methods with respect to data density. Specifically, we started by assessing the proportion of unknown relationships in the initial network. Then we removed interactions from the network until one more dyad had no interactions, which corresponds to an increase in the proportion of unknown relationships, and stored the resulting matrix. Then we repeated the removal procedure until we reached a matrix with 70% unknown dyads (which represents a fairly sparse network) and stored each intermediate matrix (Supplement S5, section 3). For example, a matrix with 10 individuals and hence 45 dyads, with all dyads initially observed, could have 31 dyads removed until reaching 70% unknown dyads. This would result in 32 matrices (31 removal steps + the initial matrix). To speed up computation, whenever the number of resulting matrices was larger than 12, we chose the initial matrix and randomly selected another 11 matrices and excluded the remaining ones from analysis.

The quantity of interest in this experiment was the relationship between steepness and the proportion of unknown relationships. For doing this, we separately fitted a linear regression for each dataset (set of up to 12 matrices) and for each steepness algorithm. The steepness value was the response variable and the proportion of unknown relationships was the predictor variable. The slope of this regression provided information on how a method responds to unknown relationships.

A perfect method would produce the same steepness measure regardless of the proportion of unknown relationships, and hence the slope would be zero, that is, a method returns (on average) the same steepness for each value of unknown relationships. If the slope is positive, the steepness value becomes larger with increasing unknown relationships, that is, less dense matrices. If the slope is negative, the steepness value becomes smaller with increasing unknown relationships. This latter pattern is what we expected for the classic steepness metrics, and we expected it to be more pronounced, that is, more negative, in the $P_{\gamma}$-based steepness compared to the $D_{\delta}$-based steepness (de Vries et al., 2006).

### 2.1 Datasets

#### 2.1.1 Artificial datasets

We generated 1000 artificial dominance interaction matrices. The overall goal was to produce interaction networks with specific steepness (determined by us), which were heterogeneous with respect to group size, total number of interactions and how interactions were distributed among individuals and dyads.

For each artificial matrix, we first set a group size, $n_{\text{group size}}$, between 5 and 25, which corresponds to the range of the majority of published dominance interaction matrices (Neumann, 2022b). Each individual was assigned a dominance rank that was an integer ranging from 1 through $n_{\text{group size}}$. Then we set the number of interactions dependent on group size with $n_{\text{interactions}} = n_{\text{group size}}^x$, where $x$ was a random number from a uniform distribution ranging between 1.8 and 2.8. For a group with five individuals, this leads to interaction numbers ranging between 19 and 91, and for the largest groups with 25 individuals to between 329 and 8208 interactions. The input steepness (our ‘ground truth’) for each matrix was set to a random number between 0.2 and 1 drawn from a uniform distribution (Supplement S5, section 2).

Next, we introduced two kinds of biases in how the number of interactions were distributed across dyads (Supplement S5, section 2). First, interaction probabilities of dyads depended on how close two individuals were in rank. This parameter ranged from all dyads having the same underlying propensity to interact regardless of rank (our low bound for steepness), it was still the higher-ranked individual. Since our data generation at its core relied on dyadic win/loss proportions, the process resembled the algorithm for classic David’s scores (Supplement S5, section 2).

In the final step, we generated interactions and their outcomes. First, we randomly assigned a frequency of interactions to each dyad that was proportionate to the dyadic interaction probabilities as defined above under the constraint that all dyadic frequencies summed to the total number of interactions defined above. Then, for each dyad, we sampled from a binomial distribution to determine the number of wins for the higher-ranked individual in that dyad. The probability for a success (i.e. a win) in the sampling process was $pr = (s + 1) / 2$, where $s$ was the desired steepness. For example, in a dyad with 10 interactions and $s = 0.8$, the higher ranked individual is expected to win 9 interactions, while for $s = 0.2$ (our low bound for steepness), it would be expected to win 6. This transformation was necessary to ensure that for steepness values below 0.5 it was still the higher-ranked individual who won (on average) more interactions than the lower-ranked individual. The number of losses of the lower-ranked individual in a dyad was then calculated as the difference between the total number of interactions in that dyad and the number of wins for the higher-ranked individual. Since our data generation at its core relied on dyadic win/loss proportions, the process resembled the algorithm for classic David’s scores (Supplement S5, section 2).

#### 2.1.2 Empirical datasets

We compiled a database that contained 978 published dominance interaction matrices. Of these, we only kept those that met the following criteria: The number of individuals was at least five, the
proportion of unknown relationships was less than 0.5 and all individuals were observed in at least one interaction. After this selection process 670 unique matrices remained (Neumann, 2022b, Supplement S1).

2.1.3 | Algorithms used

In addition to our new steepness measure, we also subjected five other algorithms/variants to our evaluations. The first were the two versions of classic steepness, based on $D_D$ and $P_p$ winning proportions as described above (de Vries et al., 2006). We refer to these two as $DS_D$ and $DS_p$.

The third steepness metric was based on a Bayesian version of David’s scores, which we also developed in the course of this study and describe in more detail in the Supplement S5, section 4. We refer to this method as $DS_{Bayes}$.

The fourth algorithm was based on repeatability of Elo-ratings (Sánchez-Tójar et al., 2018). Here an interaction matrix is translated into a large number of randomized interaction sequences (typically 1000). Each sequence is then subjected to the Elo-rating algorithm, which results in a set of multiple ratings for each individual (one rating per individual per randomized sequence). For these ratings, repeatability (also known as the intra-class correlation coefficient, Nakagawa & Schielzeth, 2010) across individuals is calculated, which serves as steepness estimate (Sánchez-Tójar et al., 2018). We refer to this method as $Elorpt$.

The final algorithm was simply the proportion of interactions that go against the rank order (upward steepness). To this end, an ordinal ranking of individuals is produced first. Here we use classic David’s score to produce this ranking. Then the interaction matrix is reordered according to the obtained ranking. The upward steepness is just the proportion of interactions below the diagonal divided by the total number of interactions. It is noteworthy to say that this index can, at least theoretically, be zero if the produced ranking is completely false, that is, the initially produced ranking is the opposite of the ‘true’ ranking and there are no entries above the diagonal in the dominance matrix.

The results of all algorithms share the same scale, that is, they all range between 0 and 1 where 0 indicates a shallow hierarchy and 1 indicates a maximally steep hierarchy. This feature allows the comparison between the different methods, at least approximately, given that the scales of the underlying dominance scores differ between the various methods. Also note that although STEER and steepness based on Bayesian David’s scores produce posterior distributions, we reverted to using posterior medians (i.e. point estimates) in the quantitative evaluations to simplify comparisons with the other methods.

2.1.4 | Software

All data were generated and analysed with the EloSteepness (Neumann, 2022a) and EloRating (Neumann & Kulik, 2020) packages, which are based on Stan (accessed through rstan and cmdstanr [Gabry & Češnovar, 2022; Stan Development Team, 2022]) and Rcpp (Eddelbuettel & Francois, 2011). Steepness based on repeatability was obtained from the aniDom package (Farine & Sánchez-Tójar, 2021). All data we used and generated as well as code to replicate our analyses are in the EloSteepness.data data package (Neumann, 2022b and Supplement S2 and S3). The EloSteepness package also contains a tutorial, illustrating how STEER can be applied (see also Supplement S5, section 5 for a brief overview).

3 | RESULTS

3.1 | Recovering ground truth

For networks with high density (less than 5% unknown relationships), all methods showed a close and, with the exception of $Elorpt$, linear relationship between the estimated steepness and the steepness we set during the data generation of our artificial networks (Figure 4). $DS_p$ steepness performed best overall, that is, it had the least bias, although it still slightly overestimated smaller ground truth values. $DS_D$ and $DS_{Bayes}$ steepness tended to underestimate ground truth, especially at high values of ground truth. $Elorpt$ overestimated steepness at high ground truth values and underestimated steepness at low ground truth values. $Elorpt$ also tended to slightly overestimate ground truth steepness. Finally, upward steepness consistently produced higher steepness values than expected, although the relationship was still a linear one.

Although actual ground truth was unknown for our empirical datasets, we nevertheless analysed these data in a similar way, defining $DS_p$ steepness as ‘ground truth’. Keeping this severe limitation in mind, our results were qualitatively similar to those presented above for the artificial data (Supplement S5, section 6).

3.2 | Handling sparse datasets/removal experiment

All steepness methods based on David’s scores, including Bayesian David’s scores, showed a strong negative dependence (as expected) on the proportion of unknown relationships (Figures 5 and 6), that is, steepness decreased the sparser a dataset became, which resulted in negative slopes (as depicted in the figures). In contrast, the two Elo-rating-based methods (including STEER) and the simple upward steepness showed on average little dependence on unknown relationships.

4 | AN APPLIED EXAMPLE

To illustrate a potential application of our new measure, we revisited the study by Balasubramaniam et al. (2012) who investigated variation in dominance hierarchy steepness in social groups belonging to different species of macaques (Macaca sp.). Our main
Objective was to partition the observed variation in steepness across macaque groups into three distinct variance components: (1) variance due to the phylogenetic relationships between species, (2) between-species variance independent of the phylogeny and (3) within-species variance. The first variance component reflected the phylogenetic relationships between species. The second component accounted for variance between species independent of the phylogeny, due to repeated measurements of groups belonging to the same species. The third component captured residual variance unaccounted for by the two first components, that is, variation within species.

In our analysis, we used data from 67 groups belonging to 13 species (Figure 7). The response variable in our model was steepness, which itself was represented by posterior distributions of the STEER estimates (estimated in the same model). In other words, we modelled steepness (from the raw interaction data) simultaneously with estimating the different variance components in one single model. To account for the phylogenetic relationships between groups, we included the n×n phylogenetic distance matrix, and the n×n within-group distance matrix in one single model.
species we assumed a Brownian motion model, using a pruned tree from a consensus phylogeny of primates (Arnold et al., 2010). For simplicity, we logit-transformed steepness to fit models with normal likelihood, rather than the arguably more appropriate beta likelihood (Nakagawa & Schielzeth, 2010; Warton & Hui, 2011). This model resembled a generalized phylogenetic linear mixed model (de Villemereuil & Nakagawa, 2014; Hadfield & Nakagawa, 2010). We coded this model in Stan (Stan Development Team, 2022; see Neumann (2022b) for data and model code, see also Supplement S5, section 7).

We found substantial variation in steepness across macaque social groups (Figure 7). Within-species variation represented the largest amount of variance explained by our model (median = 0.87, 89% CI: 0.74 – 1.02). Compared to this, between-species variance and variance due to phylogenetic relationships were substantially smaller (between-species: median = 0.27, 89% CI: 0.03 – 0.73, phylogenetic: median = 0.33, 89% CI: 0.04 – 0.86).

Although we did not formally quantify the phylogenetic signal in steepness, we found that phylogeny explained at least some of the variation found in steepness across macaque species. This result supports one key finding by Balasubramiam et al. (2012), and it is in line with a wider literature showing that many behavioural traits exhibit phylogenetic signals (Kamilar & Cooper, 2013).

In contrast to Balasubramiam et al. (2012), however, we did not find any indication that between-species variation in steepness is larger than within-species variation. Rather, we found the opposite: within-species variation was substantially larger than between-species variation (median within = 0.87 vs. median between = 0.27). The most likely explanation for this is that there were several key methodological differences between our study and that of Balasubramiam et al. (2012). First, we used our new algorithm to estimate steepness. Second, we estimated phylogenetic, between-species and within-species variation of dominance hierarchy steepness in a single model, which is arguably more appropriate than estimating the three quantities with separate analyses (Garamszegi, 2014). Finally, we also used a larger dataset, which comprised more groups and species. In combination, these points suggest that Balasubramiam et al. (2012) underestimated within-species variation. This argument is supported by one additional analysis we performed. When we fitted our model to the data originally used by Balasubramiam et al. (2012), we also found that the largest variance component reflected within-species variation, that is, we found no evidence for larger between-species variation in this analysis either, although the relative differences between the variance components were less pronounced (Supplement S5, section 7).

Overall, our results indicate that there is a lot of within-species variance in steepness left unexplained by our model. It therefore seems likely that other factors drive this variation. Whether these factors pertain to the physical environment (e.g. predation risk or resource distribution) and/or reflect the social environment (e.g. group size or kin structure) remains to be seen. Future comparative studies should therefore incorporate such factors as much as possible in order to understand what drives variation in steepness (and social behaviour more generally).

Ultimately, we do agree with Balasubramiam et al. (2012) and also others (e.g. DeCasien et al., 2017; Schino & Aureli, 2008; Thierry et al., 2008) that it is important to incorporate phylogeny when asking broader questions about what explains across-species variation in or relationships between biological traits (see also Blomberg et al., 2003; Kamilar & Cooper, 2013). As our example illustrates, we have the statistical tools available to model within-species variation in such analyses alongside adjusting for the phylogenetic relationships between species and additional between-species variation (see also de Villemereuil & Nakagawa, 2014; DeCasien et al., 2017; Garamszegi, 2014; Hadfield & Nakagawa, 2010).

5 | DISCUSSION

In this study, we presented STEER, a novel algorithm to quantify steepness of animal dominance hierarchies and evaluated its performance in comparison to other available methods. The new method based on Bayesian estimation of Elo-ratings is a considerable improvement over existing algorithms to estimate hierarchy steepness. It recovered ground truth faithfully and it showed little systematic dependence on unknown relationships. In contrast, the second method we developed, based on a Bayesian implementation of David’s scores still outperformed classic steepness with respect to precision and bias, but performed overall less well than STEER.

Note that for our evaluation we ignored the fact that our method’s output represents a Bayesian posterior distribution of steepness. Rather, we treated it here as a point estimate (the median of

![FIGURE 7 An example of applying STEER in a comparative study. Here we estimated steepness in 67 groups of 13 macaque species and assessed variance due to phylogeny, between-species variance independent of phylogeny, and within-species variance (residual variance). Within-species variance was larger than phylogenetic and between-species variance.](image-url)
the posterior distribution) to simplify the comparison with the other methods, which provide point estimates. However, it is clearly beneficial to have a sense of uncertainty of steepness, which our new method provides via credible intervals, which makes inference about uncertainty much more explicit and straightforward. For example, a 89% credible interval for a median steepness of 0.8 that ranges from 0.31 to 0.97 should be taken more cautiously than an interval of 0.73 to 0.85 for the same median steepness. These uncertainties then can either be used descriptively when characterizing the hierarchy of an animal group, or can be carried forward in cases where steepness is a predictor or response variable in analyses that comprise multiple groups within or between species (e.g. Balasubramaniam et al., 2012; Kaburu & Newton-Fisher, 2015; Schino & Aureli, 2008).

We did the latter in our example analysis (Figure 7, see also figure S10 for an analysis that uses point estimates of steepness).

Importantly, our method is currently assuming that steepness is fixed and does not change over time. This assumption is implicit if there is no sequential information available in the first place like in the networks, both empirical and artificial, that we used for our evaluations. Since the Elo-raging method that underlies STEER requires sequences, it remains to be seen whether the approach we took (using randomized sequences) introduces systematic bias in steepness assessment. To be explicit, if we estimate steepness from an aggregated network our model assumes that any temporal patterns (for example dyadic rank changes) that could be reflected in the underlying, but unknown, sequence of interactions are either absent or negligible. This assumption, whether stated or not, applies to any ranking algorithm using time-aggregated interaction data (for an illustration, see Neumann et al., 2018).

If sequential information is available we can still fit the STEER model (function elo_steepness_from_sequence()), thereby removing one source of uncertainty because randomization of the sequence is not required. Crucially, this should still be considered a ‘static’ approach though because steepness is assumed to be fixed throughout the sequence of interactions, that is, summed winning probabilities and steepness are calculated after all interactions were evaluated. However, it seems clearly beneficial to extend STEER to dynamic networks, where the sequence of interactions is known and where steepness is allowed to vary across the sequence. Then it would be possible to assess (temporal) changes in individuals’ inherent fighting ability (‘rank changes’) and temporal changes at the group level (e.g. steepness). This would allow extracting steepness alongside individual dominance scores after each interaction (or calendar date, or any desired time unit) in a longitudinal fashion. The latter is already implemented (Goffe et al., 2018), but implementing the former requires more work. For example, it could be hypothesized that steepness should be larger when competition for resources becomes more intense. When this competition is measured quantitatively (e.g. number of receptive/fertile females available to males who compete for conceptions, availability of high-quality food or territories), then a temporal implementation would allow monitoring the relationship between steepness and competition without the need of dividing datasets into arbitrary time blocks.

In line with recent versions of the Elo-rating model (Franz et al., 2015; Goffe et al., 2018), our algorithm also estimated the relevant parameters, k value and start ratings, from the interaction data rather than relying on fixed values supplied by the user (Albers & de Vries, 2001; Neumann et al., 2011). Doing so allows the uncertainty in their estimation be carried forward into the actual steepness estimation. In addition, estimating k and start ratings has the advantage that it avoids the need for setting them at arbitrary values. For example, preliminary results suggest that fixing k rather than estimating it can lead to biased steepness values (Supplement S5, section 8) and therefore should be avoided.

### 5.1 Performance of other methods

The two steepness measures based on classic David’s scores performed poorly in our evaluations. While both recovered ground truth well overall (Dq less so compared to P0), both of them were strongly affected by network density. This result is not surprising because their biases and susceptibility to underestimating steepness are well known (e.g. Klass & Cords, 2011) and provided the original impetus for the development of our new method. We therefore think it justified to suggest abandoning their use given that we now have more robust alternatives.

The steepness estimate provided by assessing repeatability of Elo-ratings recovered ground truth well, albeit with some nonlinearity, and its dependence on network density was overall comparable to STEER. We do however see two disadvantages with this metric. First, it resulted in slightly biased steepness estimates by overestimating large and underestimating low steepness values. Second, there is no clear theoretical rationale for why repeatability of individual dominance scores should mechanistically provide us with an index that describes average differences in individual dominance success, that is, steepness. Both arguments by themselves do not appear to be major issues, but they illustrate the need for future work on Elo-rating specifically on what causes this nonlinearity and more generally, whether it is an appropriate steepness estimator.

The upward index, despite its simplicity, performed surprisingly well with respect to decreasing network density, although it consistently overestimated ground truth steepness. It must be noted though that it relies on ranking the individuals in the first place, which in itself is susceptible to biases. As such, its theoretical minimum of 0 can only be reached if the ranking that is initially required is completely wrong, which seems an unlikely outcome of any ranking algorithm known to us (e.g. Bayly et al., 2006; Neumann et al., 2018). As a result, it appears that the upward index is likely to overestimate steepness, which indeed seems to be the case (Figure 4). Furthermore, one other potential drawback of this method is that it pools all interactions across dyads and hence might be susceptible to dyads that interacted disproportionately frequently.

All these options share the absence of a direct assessment of uncertainty because they are point estimates. As noted above, the
steepness derived from Bayesian Elo-rating provides such an assessment in a very explicit fashion, which in itself is a major advantage.

Lastly, we also want to point to our implementation of Bayesian David’s scores. While its performance in the steepness context was not quite up to our new measure based on Elo-rating, it still performed better than steepness based on classic David’s scores. Given the popularity of David’s scores as a means to quantify individual dominance strength (in addition to forming the basis of classic steepness), it might be a fruitful follow-up to properly validate these scores as a dominance measure in their own right.

5.2 | Guidelines/recommendations

When applying our new steepness measure, we recommend to present the results in a way that does justice to its probabilistic (Bayesian) nature, that is, provide readers with an appropriate assessment of the uncertainty associated with the outcome of the analyses. We thus recommend to provide numerical (mean, median, credible interval) as well as a graphical presentation of the posterior distribution of the steepness index.

Sánchez-Tójar et al. (2018) provided some guidelines regarding the required amount of data, that is, the average number of recorded interactions per individual, that is necessary to infer a reliable dominance hierarchy. Specifically, they suggested that at least 10 (better 20) interactions per individual need to be observed (Sánchez-Tójar et al., 2018, p. 605). We find it hard to make an analogous statement with regards to our new steepness measure. We want to avoid as much as possible recommending arbitrary cut-offs, which potentially force researchers to make binary decisions. Arguably, this could lead to under-reporting of data and results simply because a researcher who failed to reach a specific cut-off might choose to not report their results. Rather, we would advise to let the data ‘speak for themselves’. The reason for this recommendation is that we conceive of steepness as a distribution rather than a point estimate, which allows assessment (via the width of the steepness distribution) of how confident we can be in our results given our observations. Seen like this, we should expect that more observed interactions lead to narrower steepness distributions. At the same time, we should not expect that more observed interactions lead to changes in the central tendencies (median or mean) of the steepness distribution.

Figure 8 shows these two relationships between data density (number of interactions divided by number of individuals) and point estimate (median of distribution), and between data density and the width of the 89% credible interval around the point estimate (as a direct measure of uncertainty of the point estimate). This figure illustrates that the point estimate is by and large independent of data density, although it appears that very large steepness values are not likely to appear with low numbers of observed interactions. More importantly (and not surprisingly), the uncertainty decreases with increasing observations, that is, credible intervals become narrower with larger numbers of observed interactions. Neither of these two plots suggests any clear cut-off. In other words, rather than using some more or less arbitrary cut-off to decide whether the observed steepness is an accurate description of the world, we find it more intuitive to let the data speak for themselves: a wide credible interval should make us less confident than a narrow credible interval.

We do, however, agree with Sánchez-Tójar et al. (2018) in stressing the importance of reporting data characteristics of the studied networks (average number of observations per individual, proportion of unknown relationships). Ideally, this reporting is accompanied by the raw interaction data, preferably in sequence form, from which steepness was estimated.

One additional lesson from Figure 8 is that while the STEER point estimates did not reach values close to 1 with low data density, these are also the cases where the credible intervals around those point estimates are fairly wide. This illustrates again the importance of interpreting point estimates alongside the actual posterior distributions.

The final issue that needs consideration is the number of randomized sequences to use when applying our new method. In principle, a simple rule would state that one would use as many randomizations as possible. For example, many statistical tests use 1000 randomizations as a default. For our evaluations, this presented a practical problem, due to limited computational resources. We therefore established a rough guide, which uses 50 randomizations for matrices with less than 100 interactions, 20 randomizations for matrices with between 100 and 500 interactions, and 5 for matrices with more than 500 interactions. To establish this guide, we took the following approach. For each dataset, we randomly selected one matrix (which is either the full matrix, or one of the matrices from which interactions were removed). We then applied our algorithm 1000 times to each matrix.
As our ‘true’ steepness value, we took the median of all posterior samples from 500 of those randomizations. From the remaining sequences, we calculated steepness multiple times from one sequence, two sequences, 5, 10, 20, 50, 100 and 200 sequences. We then checked at which of these increments the derived steepness differed less than 0.01 from our ‘true’ value. The result of this procedure gives an approximate and rough idea at which number the estimated steepness distributions levels off, that is, at which value of randomizations do we get sufficiently close to the true value. The results of this approach are presented in Figure 9 and they indicate that generally, matrices with fewer interactions require larger numbers of randomizations than matrices with many interactions.

To be clear, this analysis just serves as a rough guide for setting the number of randomized sequences in our method evaluation. And it is important to remember that whereas our new method provides actual posterior distributions, we had to resort to using point estimates simply for being able to compare our method’s results with the results of the alternative methods. In practice, or if in doubt, it is advisable to set the number of randomizations higher than what we used here and 100 seems a fairly safe value.

FIGURE 9 Number of randomized sequences needed to achieve stable results. Matrices with fewer interactions require more randomizations than larger matrices. The shaded areas represent the rule of thumb we used in our evaluations. The majority of matrices achieved stable results when applying these cut-offs.

CONCLUSIONS

We set out to develop a method that allows unbiased and precise estimation of dominance hierarchy steepness with explicit uncertainty assessment. Our results demonstrate that STEER, steepness based on Bayesian Elo-rating via summed winning probabilities, provides such a measure. By also providing the EloSteepness R package we make this assessment as user-friendly as possible.

ACKNOWLEDGEMENTS

We thank Han de Vries and Andrew Robbins for the initial impetus to start this work, and Roger Mundry and Julie Duboscq for numerous discussions. We also thank Han de Vries, Andrew Robbins, David McDonald and Cédric Girard-Buttoz for helpful comments on earlier drafts of this manuscript. The associate editor and two reviewers provided very helpful feedback. Funding was provided by the German Research Foundation project RTG 2070 Understanding Social Relationships (project number 254142454). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest exists.

REFERENCES

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. https://doi.org/10.1006/anbe.2000.1571

Amici, F., Widdig, A., MacIntosh, A. J. J., Francés, V. B., Castellano-Navarro, A., Caicoya, A. L., Karimullah, K., Maulany, R. I., Ngakan, P. O., Hamzah, A. S., & Majolo, B. (2020). Dominance style only partially predicts differences in neophobia and social tolerance over food in four macaque species. *Scientific Reports, 10*, 22069. https://doi.org/10.1038/s41598-020-79246-6

Arnold, C., Matthews, L. J., & Nunn, C. L. (2010). The 10kTrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology, 19*, 114–118. https://doi.org/10.1002/evan.20251

Balasubramaniam, K. N., Dittmar, K., Berman, C. M., Butovskaya, M. L., Cooper, M. A., Majolo, B., Ogawa, H., Schino, G., Thierry, B., & de Waal, F. B. M. (2012). Hierarchical steepness and phylogenetic models: Phylogenetic signals in *Macaca*. *Animal Behaviour, 83*, 1207–1218. https://doi.org/10.1016/j.anbehav.2012.02.012

Barrett, L., Henzi, S. P., Weingrill, T., Lycett, J. E., & Hill, R. A. (1999). Market forces predict grooming reciprocity in female baboons.
Proceedings of the Royal Society B: Biological Sciences, 266, 665–670. https://doi.org/10.1098/rspb.1999.0687
Bayly, K. L., Evans, C. S., & Taylor, A. (2006). Measuring social structure: A comparison of eight dominance indices. Behavioural Processes, 73, 1–12. https://doi.org/10.1016/j.beproc.2006.01.011
Blomberg, S. P., Garland, T., Jr., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution, 57, 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x
Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? Using the EloChoice package to assess pairwise comparisons of perceived physical strength. PLoS ONE, 13, e0190393. https://doi.org/10.1371/journal.pone.0190393
David, H. A. (1987). Ranking from unbalanced paired-comparison data. Biometrika, 74, 432–436. https://doi.org/10.1093/biomet/74.2.432
de Villemereuil, P., & Nakagawa, S. (2014). General quantitative genetic methods for comparative biology. In L. Z. Garamszegi (Ed.), Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice (pp. 287–303). Springer. https://doi.org/10.1007/978-3-662-43550-2_11
de Vries, H., Stevens, J. M. G., & Vervaekte, H. (2006). Measuring and testing the steepness of dominance hierarchies. Animal Behaviour, 71, 585–592. https://doi.org/10.1016/j.anbehav.2005.05.015
DeCasien, A. R., Williams, S. A., & Higham, J. P. (2017). Primate brain size is predicted by diet but not sociality. Nature Ecology & Evolution, 1, 1–7. https://doi.org/10.1038/s41559-017-0112
Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ integration. Journal of Statistical Software, 40, 1–18. https://doi.org/10.18637/jss.v040.i08
Elo, A. E. (1978). The rating of chess players, past and present. Arco.
Farine, D. R., & Sánchez-Tójar, A. (2021). aniDom: Inference dominating hierarchies and estimating uncertainty. https://CRAN.R-project.org/package=aniDom
Flack, J. C., & de Waal, F. B. M. (2004). Dominance style, social power, and conflict management in macaque societies: A conceptual framework. In B. Thierry, M. Singh, & W. Kaumanns (Eds.), Macaque societies (pp. 157–182). Cambridge University Press.
Franz, M., McLean, E., Tung, J., Altmann, J., & Alberts, S. C. (2015). Self-organizing dominance hierarchies in a wild primate population. Proceedings of the Royal Society B: Biological Sciences, 282, 20151512. https://doi.org/10.1098/rspb.2015.1512
Gabry, J., & Češnovar, R. (2022). cmdstanr: R interface to CmdStan. https://mc-stan.org/cmdstanr/
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. https://doi.org/10.1016/S0003-1944(02)00226-7
Garamszegi, L. Z. (2014). Uncertainties due to within-species variation in comparative studies: Measurement errors and statistical weights. In L. Z. Garamszegi (Ed.), Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice (pp. 157–199). Springer. https://doi.org/10.1007/978-3-662-43550-2_7
Geffe, A. S., Fischer, J., & Sennhenn- Reulen, H. (2018). Bayesian inference and simulation approaches improve the assessment of Elo-ratings in the analysis of social behaviour. Methods in Ecology and Evolution, 9, 2131–2144. https://doi.org/10.1111/2041-210X.13072
Haddfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology, 23, 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x
Jennings, D. J., Gammell, M. P., Carlin, C. M., & Hayden, T. J. (2006). Is difference in body weight, antler length, age or dominance rank related to the number of fights between fallow deer (Dama dama)? Ethology, 112, 258–269. https://doi.org/10.1111/j.1439-0310.2006.01154.x
Kaburu, S. S. K., & Newton-Fisher, N. E. (2015). Egalitarian despots: Hierarchy steepness, reciprocity and the grooming-trade model in wild chimpanzees, Pan troglodytes. Animal Behaviour, 99, 61–71. https://doi.org/10.1016/j.anbehav.2014.10.018
Kamlar, J. M., & Cooper, N. (2013). Phylogenetic signal in primate behaviour, ecology and life history. Philosophical Transactions of the Royal Society B: Biological Sciences, 368, 20120341. https://doi.org/10.1098/rstb.2012.0341
Klass, K., & Cords, M. (2011). Effect of unknown relationships on linearity, steepness and rank ordering of dominance hierarchies: Simulation studies based on data from wild monkeys. Behavioural Processes, 88, 168–176. https://doi.org/10.1016/j.beproc.2011.09.003
Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. Biological Reviews, 85, 935–956. https://doi.org/10.1111/j.1469-185X.2010.01041.x
Neumann, C. (2022a). EloSteepness: Bayesian dominance hierarchy steepness via Elo rating and David’s scores. https://CRAN.R-project.org/package=EloSteepness
Neumann, C. (2022b). EloSteepness.data: Data sets for package EloSteepness. https://github.com/gobbi-os/EloSteepness.data
Neumann, C. (2022c). Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies. Zenodo, https://doi.org/10.5281/zenodo.7187039
Neumann, C., Duboscq, J., Dubuc, C., Ginting, A., Irwan, A. M., Agil, M., Widdig, A., & Engelhardt, A. (2011). Assessing dominance hierarchies: Validation and advantages of progressive evaluation with Elo-rating. Animal Behaviour, 82, 911–921. https://doi.org/10.1016/j.anbehav.2011.07.016
Neumann, C., & Kulik, L. (2020). EloRating: Animal dominance hierarchies by Elo rating. https://CRAN.R-project.org/package=EloRating
Newton-Fisher, N. E. (2017). Modeling social dominance: Elo-ratings, prior history, and the intensity of aggression. International Journal of Primatology, 38, 427–447. https://doi.org/10.1007/s10764-017-9952-2
Neumann, C., McDonald, D. B., & Shizuka, D. (2018). Dominance ranks, dominance ratings and linear hierarchies: a critique. Animal Behaviour, 144, e1–e16. https://doi.org/10.1016/j.anbehav.2018.07.012
Sánchez-Tójar, A., Schroeder, J., & Farine, D. R. (2018). A practical guide for inferring reliable dominance hierarchies and estimating their uncertainty. Journal of Animal Ecology, 87, 594–608. https://doi.org/10.1111/1365-2656.12776
Schino, G., & Aureli, F. (2008). Trade-offs in primate grooming reciprocation: Testing behavioural flexibility and correlated evolution. Biological Journal of the Linnean Society, 95, 439–446. https://doi.org/10.1111/j.1095-8312.2008.01067.x
Seyfarth, R. M. (1976). Social relationships among adult female baboons. Animal Behaviour, 24, 917–938. https://doi.org/10.1016/S0003-4727(76)80022-X
Stan Development Team. (2022). rstan: The R interface to Stan. http://mc-stan.org/
Sterck, E. H. M., Watts, D. P., & van Schaik, C. P. (1997). The evolution of female social relationships in nonhuman primates. Behavioral Ecology and Sociobiology, 41, 291–309. https://doi.org/10.1007/s002650050390
Thierry, B., Aureli, F., Nunn, C. L., Petit, O., Abbeg, C., & de Waal, F. B. M. (2008). A comparative study of conflict resolution in macaques: Insights into the nature of trait covariation. Animal Behaviour, 75, 847–860. https://doi.org/10.1016/j.anbehav.2007.07.006
Thierry, B., Iwaniuk, A. N., & Pellis, S. M. (2000). The influence of phylogeny on the social behaviour of macaques (primates: Cercopithecidae, genus Macaca). *Ethology, 106*, 713–728. [https://doi.org/10.1046/j.1439-0310.2000.00583.x](https://doi.org/10.1046/j.1439-0310.2000.00583.x)

van Schaik, C. P. (1989). The ecology of social relationships amongst female primates. In V. Standen & R. A. Foley (Eds.), *Comparative socioecology* (pp. 195–218). Blackwell.

Vehrencamp, S. L. (1983). A model for the evolution of despotic versus egalitarian societies. *Animal Behaviour, 31*, 667–682. [https://doi.org/10.1016/S0003-3472(83)80222-X](https://doi.org/10.1016/S0003-3472(83)80222-X)

Warton, D. I., & Hui, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology, 92*, 3–10. [https://doi.org/10.1890/10-0340.1](https://doi.org/10.1890/10-0340.1)

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Neumann, C., & Fischer, J. (2023). Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies. *Methods in Ecology and Evolution, 14*, 669–682. [https://doi.org/10.1111/2041-210X.14021](https://doi.org/10.1111/2041-210X.14021)