A practical guide for inferring reliable dominance hierarchies and estimating their uncertainty

Alfredo Sánchez-Tójar[a,b*], Julia Schroeder[a,b], Damien R. Farine[c,d,e*]

a. Evolutionary Biology, Max Planck Institute for Ornithology, Seewiesen, Germany

b. Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK

c. Department of Collective Behaviour, Max Planck Institute for Ornithology, Konstanz, Germany

d. Department of Biology, University of Konstanz, Germany

e. Edward Grey Institute of Field Ornithology, Department of Zoology, University of Oxford, UK

*corresponding authors:

alfredo.tojar@gmail.com; telephone: +49 (0) 8157 932363; complete postal address: Max Planck Institute for Ornithology, Eberhard-Gwinner-Straße, 82319, Seewiesen, Germany.

dfarine@orn.mpg.de; telephone: +49 (0) 7732 150145; complete postal address: Fachbereich Biologie, Universität Konstanz, Universitätsstraße 10, 78464, Konstanz, Germany.

Abstract

Many animal social structures are organized hierarchically, with dominant individuals monopolizing resources. Dominance hierarchies have received great attention from behavioural and evolutionary ecologists. As a result, there are many methods for inferring hierarchies from social interactions. Yet, there are no clear guidelines about how many observed dominance interactions (i.e. sampling effort) are necessary for inferring reliable dominance hierarchies, nor are there any established tools for quantifying their uncertainty. In this study, we simulated interactions (winners and losers) in scenarios of varying steepness (the probability that a dominant defeats a subordinate based on their difference in rank). Using these data, we (1) quantify how the number of interactions recorded and hierarchy steepness affect the performance of three methods, (2) propose an amendment that improves the performance of a popular method, and (3) suggest two easy procedures to measure uncertainty in the inferred hierarchy. First, we found that the ratio of interactions to individuals required to infer reliable hierarchies is surprisingly low, but depends on the hierarchy steepness and method used. We then show that David's score and our novel randomized Elo-rating are the two best methods, whereas the original Elo-rating and the recently described ADAGIO perform less well. Finally, we propose two simple methods to estimate uncertainty at the individual and group level. These uncertainty measures further allow to differentiate non-existent, very flat and highly uncertain hierarchies from intermediate, steep and certain hierarchies. Overall, we find that the methods for inferring dominance hierarchies are relatively robust, even when the ratio of observed interactions to individuals is as low as 10 to 20. However, we suggest that implementing simple procedures for estimating uncertainty will benefit

researchers, and quantifying the shape of the dominance hierarchies will provide new insights into the study organisms.

Keywords (10): agonistic interactions, Elo-rating, David's score, dominance, dyad, hierarchy uncertainty, sampling effort, social status, steepness.

Highlights (3 to 5 bullet points, max 85 characters including spc)

- David's score and the randomized Elo-rating perform best.
- Method performance depends on hierarchy steepness and sampling effort.
- Generally, inferring dominance hierarchies requires relatively few observations.
- The R package "aniDom" allows easy estimation of hierarchy uncertainty.
- Hierarchy uncertainty provides insights into the shape of the dominance hierarchy.

INTRODUCTION

Many animal social structures are organized hierarchically, with some individuals – the dominants – monopolizing resources and therefore presumably monopolizing fitness, too. First described in the domestic fowl (Schjelderupp-Ebbe, 1922), dominance hierarchies have received great attention from empiricists and theoreticians in behavioural and evolutionary ecology. Dominance hierarchies have widely been described in insects (Choe, 1994), fishes (Polačik & Reichard, 2009), reptiles (Bush, Quinn, Balreira, & Johnson, 2016), birds (Devost, Jones, Cauchoix, Montreuil-Spencer, & Morand-Ferron, 2016) and mammals (Majolo, Aureli, & Schino, 2012), including humans (von Rueden, Gurven, & Kaplan, 2008). Extensive theoretical efforts have been made on understanding how dominance hierarchies are formed and maintained (e.g. Dugatkin & Earley, 2004; Parker, 1974; Sasaki et al., 2016).

The importance and prevalence of dominance hierarchies in nature (reviewed in Drews, 1993) has led to the development of many methods for inferring dominance hierarchies from social interactions (reviewed in Bayly, Evans, & Taylor, 2006; Briffa et al., 2013; de Vries, 1998; Whitehead, 2008), yet, clear guidelines for inferring reliable dominance hierarchies are still missing. These methods can be classified into those estimating the rank of the individuals (i.e. an ordinal score: 1,2,…,n; e.g. I&SI: de Vries, 1998; ADAGIO: Douglas, Ngonga Ngomo, & Hohmann, 2017) and those estimating non-integer indices of success from which individuals can further be ranked if required (e.g. David's score: David, 1987; Elo-rating: Elo, 1978; Fig. 1). Contrary to ranks, indices have the advantage of allowing parametric statistical testing. Index-generating methods are further classified into those based on interaction matrices (e.g. David's score) and those based on the temporal

sequence of interactions (e.g. Elo-rating; Fig. 1). What all methods have in common is that they require researchers to record agonistic dyadic interactions among individuals as input data. Despite great research effort, there are surprisingly few clear guidelines about how many interactions are needed for inferring reliable dominance hierarchies, nor are there any established tools for quantifying the uncertainty of a dominance hierarchy (i.e. determine whether sufficient observations were made).

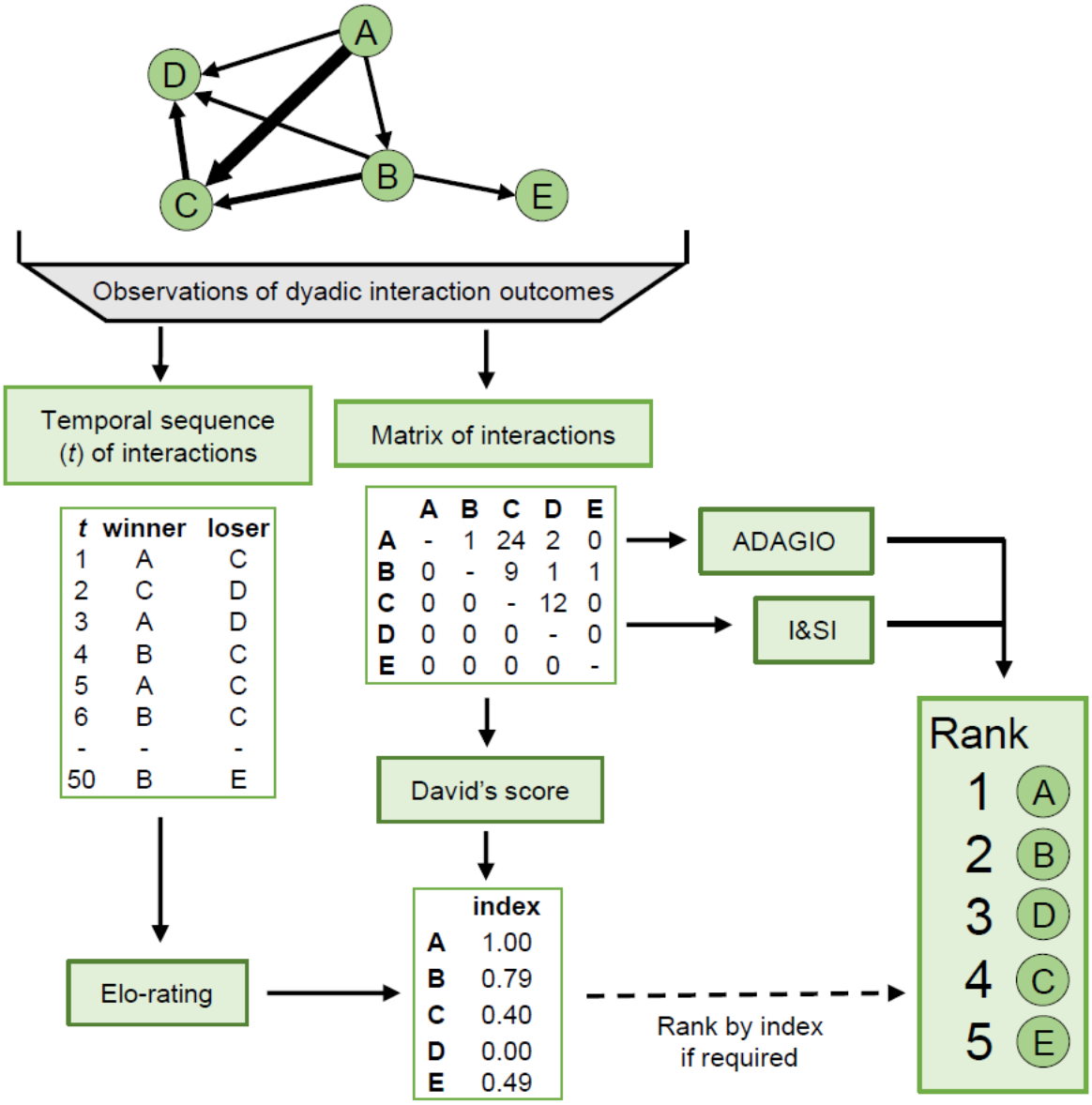


Figure 1. Diagram highlighting the different steps required to infer dominance hierarchies. First, the outcomes of dyadic agonistic interactions between individuals are recorded either in the form of a matrix or as a temporal sequence of winners and losers. Second, different methods can be used to infer either individual ranks or individual non-integer indices of success, which can further be used to rank the individuals and therefore to infer the dominance hierarchy of the group.

One source of uncertainty is the steepness of the dominance hierarchy. Dominance hierarchies can range from very steep, where dominant individuals win all conflicts, to completely flat, non-existent hierarchies, where dyadic outcomes are unpredictable. These different scenarios, which are *a priori* unknown by the researcher, are expected to affect the performance of the method inferring hierarchies. Several efforts have been made to assist researchers selecting an appropriate method, however, most of these attempts focused on comparing the level of agreement among several methods when applied to real datasets (e.g. Balasubramaniam et al., 2013; de Vries, 1998; Gammell, de Vries, Jennings, Carlin, & Hayden, 2003; Neumann et al., 2011). The problem with using real datasets is that the real steepness of the hierarchy and the real rank of the individuals are unknown. Thus, while cases where two or more methods closely match one-another could signify that they are robust, this could also mean that they suffer from a common bias, and such comparisons provide no information about their accuracy. A better approach is to (i) simulate artificial datasets containing individuals of known rank, (ii) simulate interactions among those individuals under different scenarios of known steepness, and (iii) then test the validity of the method(s) by correlating the inferred hierarchy to the original known hierarchy from step (i). This approach further allows

estimating the degree of uncertainty of the inferred ranks, and how this varies based on the steepness of the hierarchy and the number of interactions observed.

An additional source of uncertainty is the skewness in the propensity for individuals to interact. As with many other biological processes that generate count data, the number of interactions per individual often follow a Poisson distribution. This means that few individuals have many interactions, whereas most individuals have few interactions. The unequal distribution of interactions leads to interaction datasets that are sparse. Sparse datasets are very common in the dominance literature (McDonald & Shizuka, 2013) and sparseness can potentially affect the performance of the method (de Vries, 1998; Gammell et al., 2003; Neumann et al., 2011). Further, such distribution could over-inflate the perceived quality of an interaction dataset that in fact contains too few interactions to estimate most individuals' ranks.

There is increasing awareness of the need to estimate uncertainty of social data (e.g. Farine & Strandburg-Peshkin, 2015; Lusseau, Whitehead, & Gero, 2008). More than a decade ago Adams (2005) proposed a Bayesian approach to estimate hierarchy uncertainty, however, behavioural and evolutionary ecologists have not yet broadly adopted Bayesian procedures, possibly due to their apparent complexity. Thus, uncertainty is still rarely measured when studying dominance hierarchies (but see Kelstrup, Hartfelder, & Wossler, 2015; Sheppard et al., 2013). Further, estimating uncertainty is often seen as a step needed simply to 'tick a box'. Here we demonstrate that estimating uncertainty can be simple to implement, and that doing so can also generate new insights into the processes being studied.

In this study, we simulated artificial datasets and interactions under a wide range of scenarios, from very steep to completely flat, non-existent hierarchies, to: (1) quantify how the number of interactions recorded (sampling effort) and hierarchy steepness affect the performance of different methods inferring the correct hierarchy, (2) propose an amendment to a popular method, the original Elo-rating, to improve its performance, and (3) suggest two easy procedures to measure the uncertainty of the inferred hierarchies. We focus our study on three index-generating methods (plus our method) that have been commonly used in the recent literature. First, we evaluate David's score (David, 1987), a widely used matrix-like method that is based on the paired comparisons paradigm (e.g. Jennings, Carlin, & Gammell, 2009; Rat, van Dijk, Covas, & Doutrelant, 2015). Second, we evaluate the original Elo-rating (Elo, 1978), a sequential method that is becoming popular in the study of animal behaviour (e.g. Franz, Mclean, Tung, Altmann, & Alberts, 2015; Snyder-mackler et al., 2016; Strandburg-Peshkin, Farine, Couzin, & Crofoot, 2015). Following our evaluation of this method, we suggest a modification to the original Elo-rating that improves estimates and provides a measure of uncertainty for each individual's rank. Finally, we evaluate ADAGIO, a recently described method that is based on the extraction and graphical representation of directed acyclic graphs (Douglas et al., 2017). From our results, we derive recommendations on the sampling effort required to infer reliable dominance hierarchies.

MATERIALS AND METHODS

Our general approach consisted of (i) generating artificial datasets containing individuals of known rank, (ii) simulating interactions among those individuals under

different hierarchy scenarios of known steepness and propensities to interact, and (iii) testing the performance of the different methods under the different scenarios.

We implemented all of our simulations in R v.3.3.2 (R Core Team, 2016; see Sánchez-Tójar, Schroeder, & Farine, 2017). We created a user friendly R package named "aniDom" to infer dominance hierarchies using the original and the randomized Elo-rating method. The package also allows estimating hierarchy uncertainty (see below) as well as plotting the steepness of the hierarchy. We used the R package "EloRating" to calculate David's scores (Neumann & Kulik, 2014) and a separate software for the ADAGIO method (Douglas et al., 2017).

i. Generating artificial groups of individuals

We generated an artificial dataset containing 50 individuals whose ranks were sequentially assigned from 1 to 50 (i.e. linear hierarchy). Because count data (such as the number of interactions) typically follow a Poisson distribution (Zuur, Ieno, Walker, Saveliev, & Smith, 2009), we used a Poisson process to generate a varying propensity for each individual to interact. Other distributions gave qualitatively similar conclusions (Supplementary Material 1). Throughout we defined the "ratio of interactions to individuals" as the number of interactions (*d*, where one interaction is between two individuals) divided by the total number of individuals (*N*), that is $\frac{d}{N}$, rather than the arithmetic mean of the number of interactions per individual (*i*), that is $\overline{d_i}$, which could be twice the value we report. For example, in a dataset containing 10 interactions between two individuals, $\frac{d}{N}$ would equal 5, whereas $\overline{d_i}$ would equal 10. We repeated all analyses with datasets containing 10 individuals (Supplementary Material 2).

196 *ii. Simulating interactions within the group*

197 We generated simulated datasets. In each interaction dataset, the outcome of the
198 dyadic interactions was determined by the specific hierarchy scenario implemented.
199 We used a probabilistic approach to generate a wide range of hierarchy scenarios of
200 different steepness. Specifically, we modelled the expected probability of winning for
201 the higher ranked individual (*P*) following the equation:

202 $$P(dominant\ wins) = 0.5 + \frac{0.5}{1 + e^{-r\,a+b}}$$ equation 1

203 where *r* is the absolute difference in rank between the two individuals divided by the
204 maximum absolute difference in rank possible in the dataset (i.e. 50 – 1 = 49 in our
205 datasets), and thus, $0 < r \leq 1$. *a* and *b* are the values that determine the steepness of
206 the hierarchy. For example, if $b = -5$ and $a \geq 0$, the expected probability of winning for
207 the higher ranked individual is essentially 1 at all times, regardless of the difference
208 in rank between the two individuals, and thus this hierarchy is very steep (Fig. 2a).
209 By contrast, if $b = 35$ and $0 \leq a \leq 30$, the expected probability of winning for an
210 individual is essentially always 0.5, regardless of the difference in rank between the
211 two contestants, and thus there is no hierarchy (Fig. 2i). Figure 2 shows a wide
212 range of the parameter space of equation 1, i.e. it shows different possible hierarchy
213 scenarios.

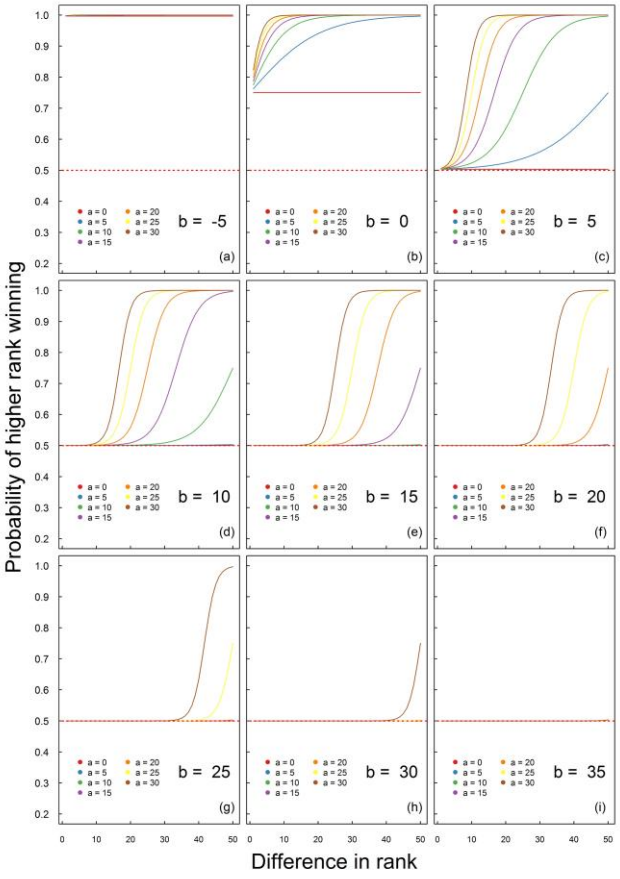


Figure 2. Parameter space of equation 1 showing a wide range of possible hierarchy scenarios that depend on the values assigned to *a* and *b*. Overall, panels are sorted from steep (panel a) to non-existent hierarchies (panel i). For each panel, steepness increases with *a*. The red dashed line shows where the probability of winning for the higher ranked individual (i.e. P(dominant wins)) equals 0.5, which corresponds to scenarios where dominance rank does not affect the probability of winning an interaction.

226 *iii. Tests*

227 *iii.a. Performance under different scenarios*

228 To quantify the combined effects of sampling effort and hierarchy steepness on the
229 ability to infer reliable dominance hierarchies, we inferred dominance hierarchies
230 from the simulated data using the original Elo-rating, and compared these to the
231 known simulated hierarchies. We explored a total of 42 hierarchy scenarios of
232 different steepness. Specifically, we examined the following *b* values of equation 1: -
233 5, 0, 5, 10, 15 and 20. For each of those *b* values, we investigated the following *a*
234 values: 0, 5, 10, 15, 20, 25 and 30. Hereafter, the initial Elo-rating for each individual
235 was set to 1 000 and *k* (a parameter in the Elo-rating function) was set to 200 (for
236 more details see Sánchez-Tójar et al. 2017). To quantify the relationship between
237 sampling effort and the performance of the method, we assessed the performance
238 for interactions datasets containing an increasing ratio of interactions to individuals
239 of: 1, 4, 7, 10, 15, 20, 30, 40 and 50. For each scenario and ratio of interactions to
240 individuals, we simulated 100 independent datasets, calculated individual ranks for
241 each dataset following the original Elo-rating and obtained the Spearman rank
242 correlation coefficient (hereafter r_S) between the inferred and the known hierarchy.
243 Therefore, if the hierarchies were identical, r_S would equal 1.

244 *iii.b. Comparing methods*

245 We evaluated the performance of the four methods (David's score, original Elo-
246 rating, randomized Elo-rating and ADAGIO) under three hierarchy scenarios of
247 intermediate steepness from equation 1, where *b* = 5, and *a* = 15, 10 and 5 (see
248 Supplementary Material 3 for other scenarios). To quantify the relationship between
249 sampling effort and the performance of the four methods, we assessed interactions

datasets containing an increasing ratio of interactions to individuals of: 1, 4, 7, 10, 15, 20, 30, 40 and 50. For each scenario and ratio of interactions to individuals, we simulated 100 independent datasets, calculated individual ranks for each dataset following each of the four methods and, for each method, obtained the r_S between the inferred and the known hierarchy.

iii.c. Randomized Elo-rating

One of the aims of the original Elo-rating is to enable tracking dynamic changes in rank over time. This means that the sequence in which interactions occur affects the inferred ranks. However, most behavioural studies assume that individual dominance rank is relatively stable over time (e.g. Poisbleau, Guillon, & Fritz, 2010). We propose an improvement of the original Elo-rating based on randomizing the order in which interactions occurred (n = 1 000 randomizations throughout). Randomising the order that interactions are recorded produces slightly different individual Elo-ratings each time, from which we can calculate a mean individual rank. This method also allows estimating the 95% range of individual ranks when run on a single interaction dataset. Note that in our method, we randomised all of the sequences of interactions because our data were simulated. However, in real datasets, one could test and account for underlying changes by randomising within certain periods of time, for example within each month, or test for winner-loser effects by randomising days but maintaining the order within each day and comparing these to full randomisations. Hereafter this new method is referred to as “randomized Elo-rating”, whereas its predecessor is referred to as “original Elo-rating”.

iii.d. Estimating hierarchy uncertainty

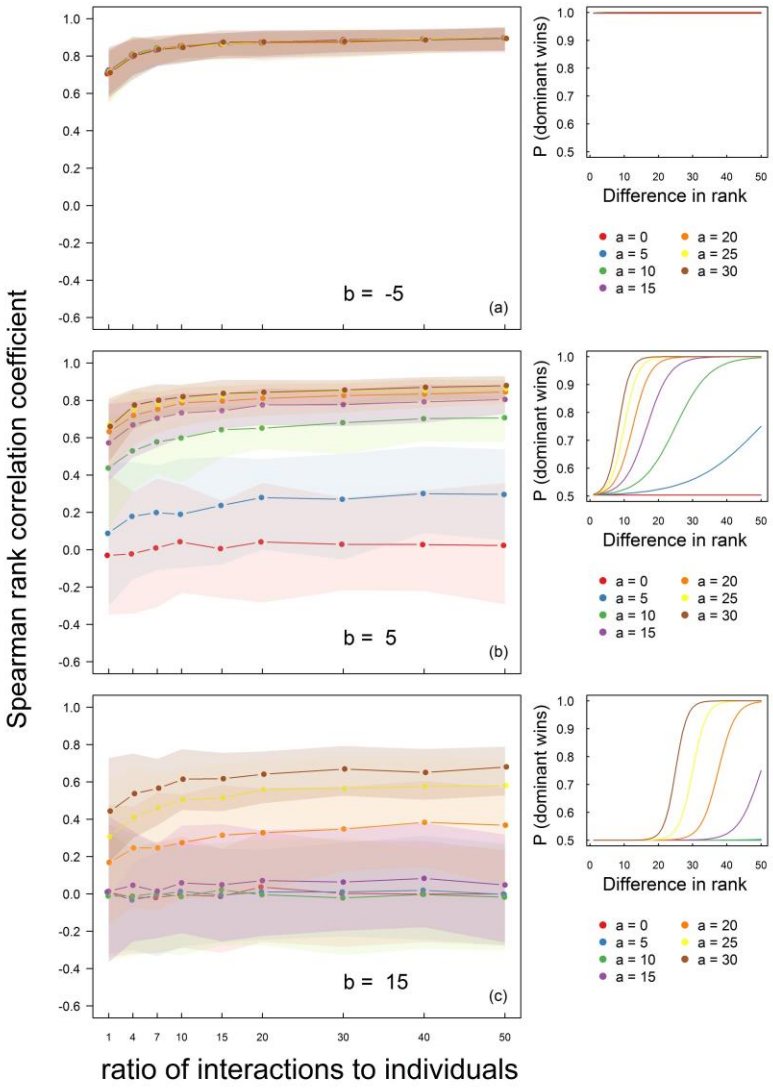
Using the randomizing Elo-rating, one can further estimate the repeatability of the *n* individual Elo-ratings. We explored repeatability using the same three scenarios described in section iii.c (see Supplementary Material 3 for other scenarios). Again, for each scenario and ratio of interactions to individuals, we simulated 100 independent interaction datasets, and calculated 1 000 individual Elo-ratings for each dataset following the randomized Elo-rating. We then calculated the repeatability of the individual Elo-ratings using the function rptGaussian() from the package 'rptR' 0.9.1.9000 (Schielzeth, Stoffel, & Nakagawa, 2016). We calculated repeatability based on Elo-ratings instead of ranks because, contrary to ranks, Elo-ratings approximately follow a Gaussian distribution.

Additionally, we tested another easy procedure to estimate hierarchy uncertainty which consists on splitting a dataset containing interactions into two halves, and then estimating the Spearman rank correlation coefficient between the two halves. We again investigated the same three hierarchy scenarios described in section iii.c (see Supplementary Material 3 for other scenarios). For each scenario and ratio of interactions to individuals, we simulated 100 independent interaction datasets, split each dataset in two, computed 1 000 individual ranks for each halve using the randomized Elo-rating, and calculated the r_S between the two inferred hierarchies.

RESULTS

Performance under different scenarios and sampling effort

We explored whether hierarchy steepness affects the performance of the method using the original Elo-rating. The performance of all four methods increased logarithmically with the ratio of interactions to individuals (Fig. 3 and 4). In most cases, the performance of the method reaches an asymptote after relatively few interactions. For very steep hierarchies, the original Elo-rating only needed a small ratio of interactions to individuals (10 or less) to satisfactorily infer the original hierarchy (e.g. Fig. 3a). In contrast, for non-existent hierarchies, the original Elo-rating visibly failed inferring the original hierarchy (e.g. $a \leq 10$ in Fig. 3c). Finally, for intermediate levels of steepness, the original Elo-rating had difficulties inferring the original hierarchy and needed a larger ratio of interactions to individuals to infer it reliably (e.g. $a = 10$ in Fig. 3b). However, in such intermediate scenarios, the improvements in the inferred hierarchy were only marginal beyond a ratio of interactions to individuals of 20.



313

314 **Figure 3**. The performance of the original Elo-rating increases with the ratio of
315 interactions to individuals and the steepness of the hierarchy. Solid lines and dots
316 represent the mean Spearman rank correlation coefficient (r_S) between the original
317 and the inferred hierarchy; shading shows the 2.5% and 97.5% quartiles. The
318 reduced right-hand side panels show the specific set of hierarchies simulated for
319 generating the interaction datasets. Overall, panels are sorted from very steep (panel
320 a) to flat hierarchies (panel c), and, for each panel, steepness increases with *a*. The
321 different hierarchy scenarios shown were created following equation 1.

The steepness of the hierarchy not only affected the amount of data required to infer reliable dominance hierarchies, but also the overall ability to do so. In general, the method performed well even when closely-ranked individuals both often win contests. For example, even when the probability of the higher ranked individual winning was only ca 0.55 for a difference in rank of 10 (Fig. 3b, a = 15), the r_s between the original and the inferred hierarchy reached up to ca 0.70 when a ratio of interactions to individuals of 20 was recorded.

Comparing methods

Next, we compared the performance of the four methods. For steep hierarchies, all four methods inferred similarly and satisfactorily the original hierarchy (Supplementary Material 3, Fig. S1a-c). Not surprisingly, for very flat hierarchies, the performance of all methods was extremely poor, even when a high ratio of interactions to individuals was recorded (Fig. 4c). At intermediate levels of steepness, we found that while all four methods had difficulties inferring the original hierarchy, the methods differed markedly in performance (Fig. 4a,b). Overall, David's score performed best, closely followed by the randomized Elo-rating. By contrast, the original Elo-rating performed relatively poorly, which is probably because the sequence of the interactions, though it has little biological relevance, can have a large impact on the resulting ranks. The recently described ADAGIO did not perform well based on our simulated winner-loser data. Further, the shapes of the curves (when adding more interactions) across the methods were quite consistent, highlighting that it is unlikely that there are scenarios where their relative performances are flipped.

The results of these simulations can also provide guidance on the number of interactions one should collect to reliably infer dominance hierarchies. Using an r_S threshold of 0.70, we suggest to record a ratio of interactions to individuals of 10 to 20 when the real steepness of the hierarchy is unknown – and either David's score or the randomized Elo-rating should be chosen.

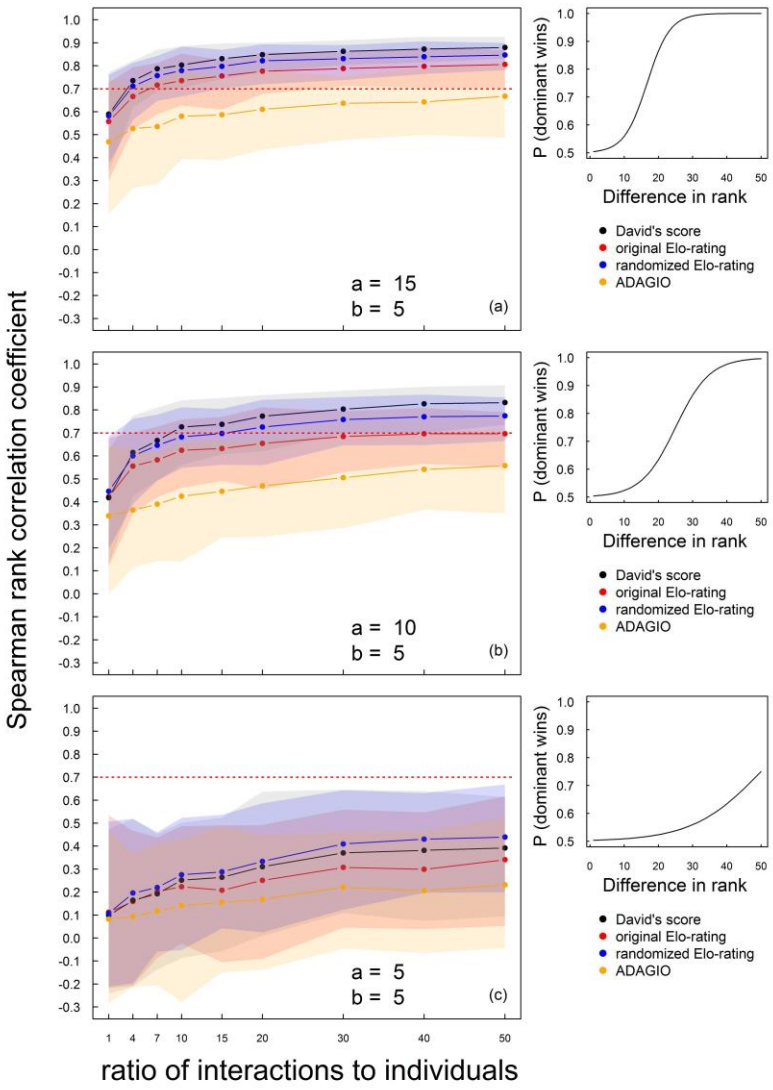


Figure 4. David's score and the randomized Elo-rating are the two best methods to infer reliable dominance hierarchies. Solid lines and dots represent the mean Spearman rank correlation coefficient (r_S) between the original and the inferred hierarchy; shading shows the 2.5% and 97.5% quartiles. The red dashed line shows

355 the suggested r_S threshold above which inferred hierarchies are highly reliable. The
356 reduced right-hand side panels show the specific hierarchy simulated for generating
357 the interaction datasets. Overall, panels are sorted from intermediate (panels a and
358 b) to very flat hierarchies (panel c). The different hierarchy scenarios shown were
359 created following equation 1.

360

361 *Hierarchy uncertainty*

362 Last, we explored two easy procedures to estimate hierarchy uncertainty. We first
363 estimated Elo-rating repeatability (Nakagawa & Schielzeth, 2010) using 1 000
364 individual Elo-ratings obtained from the randomized Elo-rating procedure. We found
365 that, for steep hierarchies, randomized Elo-rating repeatability was high (>0.90 in all
366 cases) and remained relatively stable with the ratio of interactions to individuals
367 (Supplementary Material 3, Fig. S2a-c). For intermediate levels of steepness,
368 randomized Elo-rating repeatability ranged between 0.70 and 0.95 and also
369 remained relatively stable independent of the ratio of interactions to individuals (Fig.
370 5a,b). In contrast, for very flat hierarchies, randomized Elo-rating repeatability was
371 low and decreased with the ratio of interactions to individuals (Fig. 5c). We therefore
372 suggest a repeatability threshold of 0.70 to differentiate from non-existent/very flat to
373 intermediate/steep hierarchies when enough sampling effort has been done (see
374 method 2 below). Further, our simulations also suggest that the interpretation of the
375 repeatability could include a subsampling routine to determine if repeatability is
376 stable as more data are added (i.e. intermediate/steep hierarchy, Fig. 5a,b) or
377 decreasing (i.e. non-existent/very flat hierarchy, Fig. 5c). That is, the repeatability

values provide insights into the steepness of the hierarchy (where higher repeatability scores equate a steeper hierarchy).

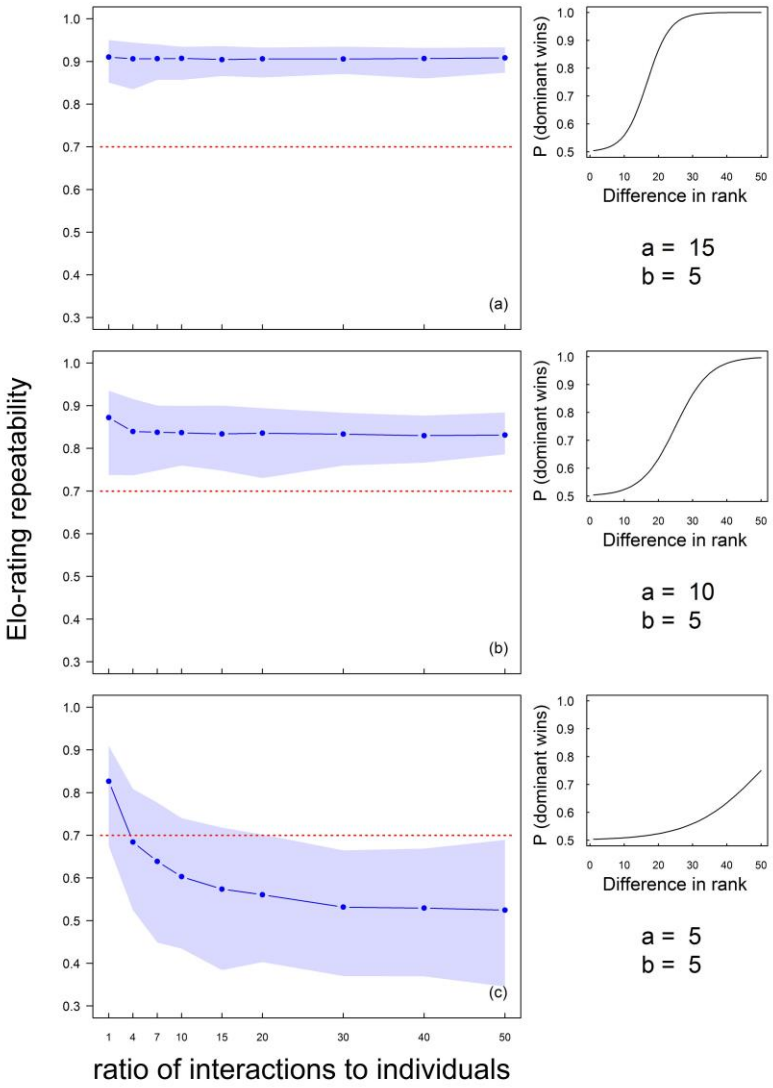


Figure 5. Randomized Elo-rating repeatability increases with the steepness of the hierarchy. Solid lines and dots represent the mean repeatability of the 1 000 individual randomized Elo-ratings estimated for each interaction dataset using the randomized Elo-rating method; shading shows the 2.5% and 97.5% quartiles. The red dashed line shows the suggested repeatability threshold above which inferred dominance hierarchies are highly reliable. The reduced right-hand side panels show the specific hierarchy simulated for generating the interaction datasets. Overall,

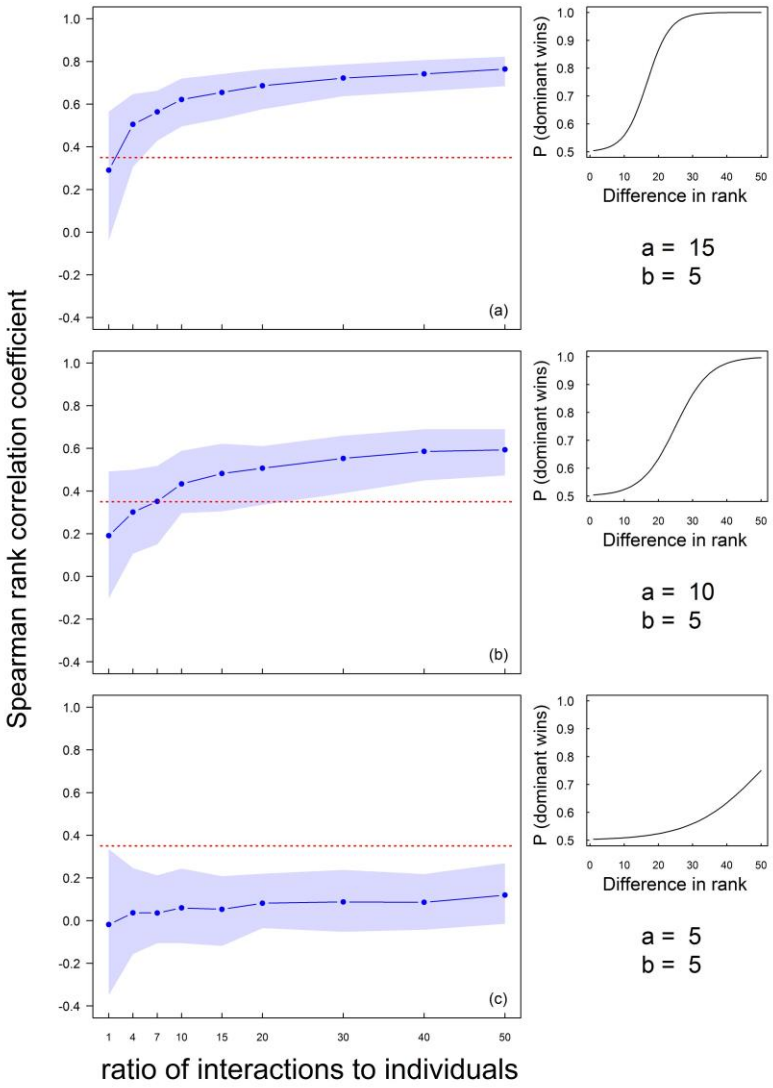
388 panels are sorted from intermediate (panel a and b) to very flat hierarchies (panel c).
389 The different hierarchy scenarios shown were created following equation 1.

390

391 We propose a second procedure to estimate uncertainty that provides useful
392 information about sampling effort. This method consists of splitting the interaction
393 dataset into two halves and estimating the agreement between the two halves. The
394 Spearman's rank correlation (r_S) between the two halves followed a similar
395 logarithmic pattern to the randomized Elo-rating performance, thus also allowing us
396 to make predictions on the level of uncertainty of our measurements and the
397 steepness of the latent hierarchy (Fig. 6). Our results show that, for very steep
398 hierarchies, the r_S quickly increased with the ratio of interactions to individuals and
399 stabilized around 0.80 (Supplementary Material 3, Fig. S3a-c). For intermediate
400 hierarchies, however, the r_S also increased but did not reach values larger than 0.80
401 (Fig. 5a,b). For very flat hierarchies, the r_S stayed relatively stable and below a
402 threshold value of 0.35 (Fig. 6c). For observational data, adding more interactions
403 into the analysis, and comparing two halves of the interaction dataset, is therefore an
404 informative method for determining whether more sampling effort is required.

405 We conclude that the higher both the randomized Elo-rating repeatability and
406 the r_S between the two halves, the steeper the inferred dominance hierarchy is. We
407 further suggest that a repeatability threshold of 0.70 and an r_S of 0.35 can be used to
408 differentiate between non-existent/very flat (highly uncertain) and intermediate/steep
409 (certain) hierarchies. Finally, by subsampling the observed data (i.e. recalculating r_S
410 with increasingly more data as we have done in Fig. 6) and observing the shape of
411 the resulting metrics is a useful way of determining if the population has been

412 adequately sampled (i.e. adding more data results in very small changes in the r_S
413 value).



414

415 **Figure 6**. The agreement between the two halves of the interaction dataset
416 increases with the ratio of interactions to individuals and the steepness of the
417 hierarchy. Solid lines and dots represent the mean Spearman rank correlation
418 coefficient (r_S) between the two halves of the datasets; shading shows the 2.5% and
419 97.5% quartiles. The red dashed line shows the suggested r_S threshold (r_S = 0.35)
420 above which inferred dominance hierarchies are highly reliable. The reduced right-
421 hand side panels show the specific hierarchy simulated to generate the interaction

422 datasets. Overall, panels are sorted from intermediate (panel a and b) to very flat
423 hierarchies (panel c). The different hierarchy scenarios shown were created following
424 equation 1.

425

426 **DISCUSSION**

427 We tested the performance of four methods to infer dominance hierarchies from
428 dyadic interactions: David's score (David, 1987), original Elo-rating (Elo, 1978),
429 randomized Elo-rating (this study) and ADAGIO (Douglas et al., 2017). We found
430 that the performance of all methods increases with the steepness of the hierarchy
431 and the ratio of interactions to individuals recorded. We showed that David's score
432 and the randomized Elo-rating are the two best methods, particularly when
433 hierarchies are not extremely steep. Further, we described two methods for
434 estimating uncertainty in a dataset of observed interactions, and found that
435 uncertainty changes predictably with hierarchy steepness and sampling effort. Based
436 on these results, we provide behavioural and evolutionary ecologists with useful
437 guidelines on the sampling effort necessary to achieve meaningful rank estimations,
438 on how to estimate the uncertainty of their results, and on how to gain insights into
439 the steepness of their hierarchy.

440 Many methods exist to infer dominance hierarchies from dyadic interaction
441 datasets, and several studies have aimed to assist researchers in choosing the
442 appropriate method (see Introduction). Contrary to most studies, we used simulated
443 interaction datasets rather than real datasets. One of the shortcomings of using real
444 datasets is that the real, latent hierarchy (if existent) is *a priori* unknown and thus, the
445 performance of a method can only be studied relative to other methods, but its

reliability cannot be assessed. For example, if all methods suffered the same inherent bias, they could reach similar results but still not reliably infer the real hierarchy. An additional advantage of simulating data is the possibility of generating an incredibly large amount of interaction datasets for each scenario studied, and thus to identify common patterns and gain more general insights on each method. By simulating interaction datasets, we showed that the four methods similarly inferred reliable hierarchies for scenarios of steep hierarchies. In contrast, the methods differed in performance when hierarchies were intermediate or very flat. We showed that David's score was the best method for intermediate hierarchies, closely followed by the randomized Elo-rating, which also outcompeted its predecessor, the original Elo-rating. Furthermore, as expected, inferred hierarchies were less reliable when using a Poisson process that generates relatively sparse datasets than when using a uniform distribution (Supplementary Material 1, section 1). Contrary to Neumann et al. (2011), our results showed that David's score still performs better than the original Elo-rating, even when data is relatively sparse. Furthermore, we found that ADAGIO's performance was sub-optimal. ADAGIO was recently shown to perform better than David's score and the original Elo-rating (Douglas et al., 2017). The main difference between the study of Douglas et al. (2017) and ours is that the former used Euclidean distances to measure the degree of agreement between original and inferred hierarchies, whereas we used r_S. While Euclidian distances (i.e. the difference between the real and estimated score) could be a reasonable measure for estimating accuracy in nonlinear hierarchies (Douglas et al., 2017), we tend to prefer relying on rank correlations. One reason for this is that a large population where all individuals have no dominance ranks (e.g. they are all 0) would yield very low

Euclidian distances that would suggest very accurate scores despite providing little information in terms of useful dominance data.

Not surprisingly, the performance of all methods increases with hierarchy steepness and the number of interactions recorded. Very steep hierarchies were reliably inferred with a ratio of interactions to individuals of only four (Supplementary Material 3, S1a-c). In contrast, intermediate hierarchies needed a greater sampling effort for inferring reliable dominance hierarchies (Fig. 4). Previous studies already indicated that the social structure of the group studied could affect the method of choice (e.g. Balasubramaniam et al., 2013; Bayly et al., 2006). Here we show the considerable effect that the steepness of the hierarchy has on the performance of the methods studied. Furthermore, many studies have commented on the importance of recording sufficient interactions to infer reliable dominance hierarchies (e.g. Gammell et al., 2003; Neumann et al., 2011). Surprisingly, no clear guidelines existed regarding the sampling effort necessary to infer reliable hierarchies from animals, which in turn has led researchers to apply either untested thresholds to define “sufficient data” (Cole & Quinn, 2011; Devost et al., 2016; Dingemanse & de Goede, 2004; Rat et al., 2015) or no threshold at all (e.g. Hauver, Hirsch, Prange, Dubach, & Gehrt, 2013; Kaburu & Newton-Fisher, 2015). Some studies even failed to report the number of interactions recorded (e.g. Campos & Fedigan, 2013; Flies, Mans, Flies, Grant, & Holekamp, 2016; Stewart & Greives, 2016), impeding researchers to assess the reliability of their results. We suggest that, unless hierarchy is *a priori* known to be very steep, researchers should aim to record a minimum ratio of interactions to individuals of 10 (or ideally 20) to ensure that the dominance hierarchy is reliably inferred. A similar number of interactions was suggested for rating chess players (Glickman & Doan, 2016) but it is considerably

larger than other suggestions for animal behaviour (Albers & de Vries, 2001). We acknowledge that recording a ratio of interactions to individuals of 10 to 20 might be challenging for species with low interaction rates such as the red deer (*Cervus elaphus*; Clutton-Brock et al., 1979), but researchers need to be aware of the potential problems of not achieving this threshold, increase (and report) sampling effort whenever possible, and ideally estimate the uncertainty of their dominance data.

Minimizing and estimating measurement error is highly recommended for the study of animal behaviour (Bradshaw, Sims, & Hays, 2007; Martin & Bateson, 2007). Indeed, there is increasing awareness of the need to estimate uncertainty of social data (e.g. Farine & Strandburg-Peshkin, 2015; Lusseau, Whitehead, & Gero, 2008), also, in the study of dominance hierarchies (Adams, 2005). Yet, uncertainty is seldom measured when analysing dominance hierarchies (but see, e.g., Kelstrup, Hartfelder, & Wossler, 2015; Sheppard et al., 2013), possibly because there are no easy-to-use tools to quantify it. Here, we provide two user friendly methods to estimate the level of uncertainty of the inferred hierarchy, and an R package to perform these. First, the randomized Elo-rating method allows calculating individual repeatability as a measure of uncertainty. We have shown that repeatability estimated by randomizing Elo-rating is a good indicator of the steepness of the latent hierarchy, and therefore of the uncertainty of the inferred hierarchy. Furthermore, it is relatively independent of sampling effort, so that, the higher the individual randomized Elo-rating repeatability, both the steeper the latent hierarchy and the more reliable the inferred hierarchy is. We suggest a repeatability threshold of 0.70 to differentiate from non-existent/very flat (highly uncertain) to intermediate/steep (certain) hierarchies. Second, we proposed that uncertainty can also be measured

by dividing the interaction dataset in two halves and calculating the level of agreement between the two. Uncertainty measured this way follows a similar pattern to that of method performance, i.e. it decreases with hierarchy steepness and sampling effort. We suggest an r_S threshold of 0.35 to differentiate from non-existent/very flat (highly uncertain) to intermediate/steep hierarchies (certain). We conclude that repeatability values and r_S below 0.70 and 0.35, respectively, are likely good indicators for a lack of a latent linear dominance hierarchy in the study system.

Additionally, we provide an improvement to the widely accepted original Elo-rating (Elo, 1978). The original Elo-rating is an sequential method proposed for rating chess players that offers some interesting features for the study of animal dominance hierarchies (Albers & de Vries, 2001; Neumann et al., 2011). We show that randomizing 1 000 times the order in which interactions occurred (randomized Elo-rating), and estimating mean individual ranks increases performance compared to the original Elo-rating, particularly when hierarchies are not extremely steep. An important feature of the original Elo-rating is that it allows visualizing hierarchy dynamics (Neumann et al., 2011), this feature is not possible when using matrix-like methods such as David's score. Visualizing hierarchy dynamics might be important to study social dynamics (Neumann et al., 2011) and/or winner-loser effects (Hsu & Wolf, 1999). Nonetheless, the randomized Elo-rating is useful when researchers have a good reason to believe that the dynamics are relatively stable and rank an inherent property of the individual. Further, we suggest that researchers can control for factors such as changing ranks or winner-loser effects by controlling which parts of the data are randomised. For example, researchers interested in tracking how individual rank changes across days, could increase the precision of the daily rank estimates by randomizing observations within each day. The randomized Elo-rating

545 further allows easy estimation of uncertainty for individual Elo-ratings or ranks. For
546 example, as discussed above, researches can estimate Elo-rating repeatability,
547 however, they can also obtain common measures of dispersion such as confidence
548 intervals or standard deviation for each individual. So far, this was only possible
549 using the Bayesian procedure proposed by Adams (2005). Finally, we believe that
550 some researchers avoid the Elo-ratings if they have recorded their data in a matrix
551 format (and thus do not have the sequence of the interactions). We suggest that in
552 these cases, the randomized Elo-rating would avoid the potential spurious results
553 that could arise by assigning a single random order to the observations.

554 Finally, we acknowledge and discuss some of the potential limitations of this
555 study. First, due to computational limitations, we only investigated four methods. We
556 however provided an R package and the R code used in this study to aid
557 researchers interested in exploring other methods (see Sánchez-Tójar et al., 2017).
558 Second, one of the limitations of the randomized Elo-rating is that it does not include
559 tied or undecided interactions. Undecided interactions are very rare (1.8% of all
560 interactions from 40 empirical interactions datasets; McDonald & Shizuka, 2013),
561 and researchers not always agree on how to interpret them (e.g. Balasubramaniam
562 et al., 2013). In fact, it could be argued that part of the undecided interactions are
563 just so because of the difficulties of understanding animal behaviour. We therefore
564 suggest that undecided interactions should always be reported but not necessarily
565 used when inferring dominance hierarchies. Third, in this study we focused on
566 inferring dominance hierarchies, i.e. we focus on ranks. However, David's score, and
567 the original and randomized Elo-ratings were originally developed for estimating
568 individual indices of (fighting) success. We believe that our suggestions can further

help researchers aiming to study individual (fighting) success rather than dominance hierarchies.

Conclusions

We have shown how sampling effort and the steepness of the underlying hierarchy (a latent feature *a priori* unknown by the researcher) affect method performance. We have suggested and provided with a new method, the randomized Elo-rating (R package "aniDom": Farine & Sánchez-Tójar, 2017). We have shown that David's score and the randomized Elo-rating are the two best methods to infer linear dominance hierarchies, particularly when the latent hierarchy is not extremely steep. Furthermore, we have introduced two easy procedures to evaluate hierarchy uncertainty at both the individual and the group level. Last but not least, we have provided clear guidelines on how much sampling effort is required to infer reliable hierarchies. We believe that the procedures outlined here are simple to implement and that the guidelines we provide will help researchers aiming to study dominance hierarchies. Finally, we hope that this work will help mitigating some of the problems recently raised (Nakagawa & Parker, 2015) in the broader field of behavioural research.

ACKNOWLEDGEMENTS

AST is member of the International Max Planck Research School (IMPRS) for Organismal Biology. We thank Antje Girndt for constructive feedback on the manuscript. AST and JS were funded by the Volkswagen Foundation. The authors declare that they have no conflict of interest.

RESOURCES

We provide all of the source code used for our data (Sánchez-Tójar, Schroeder, & Farine, 2017). We also provide with a free R package to run our implementations ("aniDom": Farine & Sánchez-Tójar, 2017). Further, we note that our implementation of the original non-randomized Elo-rating outperforms existing R packages for some scenarios (Sánchez-Tójar, Schroeder, & Farine, 2017).

REFERENCES

Adams, E. S. (2005). Bayesian analysis of linear dominance hierarchies. *Animal Behaviour*, *69*(5), 1191–1201. http://doi.org/10.1016/j.anbehav.2004.08.011

Albers, P. C. H., & de Vries, H. (2001). Elo-rating as a tool in the sequential estimation of dominance strengths. *Animal Behaviour*, *61*(2), 489–495. http://doi.org/10.1006/anbe.2000.1571

Balasubramaniam, K. N., Berman, C. M., Marco, A. De, Dittmar, K., Majolo, B., Ogawa, H., … de Vries, H. (2013). Consistency of dominance rank order: A comparison of David's scores with I&SI and Bayesian methods in Macaques. *American Journal of Primatology*, *75*(9), 959–971. http://doi.org/10.1002/ajp.22160

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

Bradshaw, C. J. A., Sims, D. W., & Hays, G. C. (2007). Measurement error causes scale-dependent threshold erosion of biological signals in animal movement

616 data. *Ecological Applications*, *17*(2), 628–638. http://doi.org/10.1890/06-0964

617 Briffa, M., Hardy, I. C. W., Gammell, M. P., Jennings, D. J., Clarke, D. D., &
618 Goubault, M. (2013). Analysis of animal contest data. In I. C. W. Hardy & M.
619 Briffa (Eds.), *Animal Contests* (p. 379). Cambridge: Cambridge University Press.
620 http://doi.org/10.1017/CBO9781139051248

621 Bush, J. M., Quinn, M. M., Balreira, E. C., & Johnson, M. A. (2016). How do lizards
622 determine dominance? Applying ranking algorithms to animal social behaviour.
623 *Animal Behaviour*, *118*, 65–74. http://doi.org/10.1016/j.anbehav.2016.04.026

624 Campos, F. A., & Fedigan, L. M. (2013). Urine-washing in white-faced capuchins: A
625 new look at an old puzzle. *Behaviour*, *150*(7), 763–798. Retrieved from
626 http://booksandjournals.brillonline.com/content/journals/10.1163/1568539x-
627 00003080

628 Choe, J. C. (1994). Sexual selection and mating system in Zorotypus gurneyi Choe
629 (Insecta: Zoraptera): I. Dominance Hierarchy and Mating Success. *Behavioral*
630 *Ecology and Sociobiology*, *34*(2), 87–93. Retrieved from
631 http://www.jstor.org/stable/4600918

632 Clutton-Brock, T. H., Albon, S. D., Gibson, R. M., & Guinness, F. E. (1979). The
633 logical stag: Adaptive aspects of fighting in red deer (Cervus elaphus L.). *Animal*
634 *Behaviour*, *27*(1), 211–225. http://doi.org/10.1016/0003-3472(79)90141-6

635 Cole, E. F., & Quinn, J. L. (2011). Personality and problem-solving performance
636 explain competitive ability in the wild. *Proceedings of the Royal Society B*, *279*,
637 1168–1175. http://doi.org/10.1098/rspb.2011.1539

638 David, H. A. (1987). Ranking from unbalanced paired-comparison data. *Biometrika*,

639 *74*(2), 432–436. http://doi.org/10.1093/biomet/74.2.432

640 de Vries, H. (1998). Finding a dominance order most consistent with a linear
641 hierarchy: a new procedure and review. *Animal Behaviour*, *55*(4), 827–843.
642 http://doi.org/10.1006/anbe.1997.0708

643 Devost, I., Jones, T. B., Cauchoix, M., Montreuil-Spencer, C., & Morand-Ferron, J.
644 (2016). Personality does not predict social dominance in wild groups of black-
645 capped chickadees. *Animal Behaviour*, *122*, 67–76.
646 http://doi.org/10.1016/j.anbehav.2016.10.001

647 Dingemanse, N. J., & de Goede, P. (2004). The relation between dominance and
648 exploratory behavior is context-dependent in wild great tits. *Behavioral Ecology*,
649 *15*(6), 1023–1030. http://doi.org/10.1093/beheco/arh115

650 Douglas, P. H., Ngonga Ngomo, A.-C., & Hohmann, G. (2017). A novel approach for
651 dominance assessment in gregarious species: ADAGIO. *Animal Behaviour*, *123*,
652 21–32. http://doi.org/10.1016/j.anbehav.2016.10.014

653 Drews, C. (1993). The concept and definition of dominance in animal behaviour.
654 *Behaviour*, *125*(3), 283–313. http://doi.org/10.1163/156853993X00290

655 Dugatkin, L. A., & Earley, R. L. (2004). Individual recognition, dominance hierarchies
656 and winner and loser effects. *Proceedings of the Royal Society B*, *271*(1547),
657 1537–1540. http://doi.org/10.1098/rspb.2004.2777

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

660 Farine, D. R., & Sánchez-Tójar, A. (2017). aniDom: Inferring Dominance Hierarchies
661 and Estimating Uncertainty. Retrieved from https://cran.r-

662 project.org/package=aniDom

663 Farine, D. R., & Strandburg-Peshkin, A. (2015). Estimating uncertainty and reliability
664 of social network data using Bayesian inference. *Royal Society Open Science*,
665 *2*, 150367. http://doi.org/10.1098/rsos.150367

666 Flies, A. S., Mansfield, L. S., Flies, E. J., Grant, C. K., & Holekamp, K. E. (2016).
667 Socioecological predictors of immune defences in wild spotted hyenas.
668 *Functional Ecology*, *30*(9), 1549–1557. http://doi.org/10.1111/1365-2435.12638

669 Franz, M., Mclean, E., Tung, J., Altmann, J., & Alberts, S. C. (2015). Self-organizing
670 dominance hierarchies in a wild primate population. *Proceedings of the Royal*
671 *Society B*, *282*(1814), 20151512. http://doi.org/10.1098/rspb.2015.1512

672 Gammell, M. P., de Vries, H., Jennings, D. J., Carlin, C. M., & Hayden, T. J. (2003).
673 David's score: a more appropriate dominance ranking method than Clutton-
674 Brock et al.'s index. *Animal Behaviour*, *66*(3), 601–605.
675 http://doi.org/10.1006/anbe.2003.2226

676 Glickman, M. E., & Doan, T. (2016). The USCF Rating System. Retrieved November
677 15, 2016, from http://www.glicko.net/ratings/rating.system.pdf

678 Hauver, S., Hirsch, B. T., Prange, S., Dubach, J., & Gehrt, S. D. (2013). Age, but not
679 sex or genetic relatedness, shapes raccoon dominance patterns. *Ethology*,
680 *119*(9), 769–778. http://doi.org/10.1111/eth.12118

681 Hsu, Y., & Wolf, L. L. (1999). The winner and loser effect: Integrating multiple
682 experiences. *Animal Behaviour*, *57*(4), 903–910.
683 http://doi.org/10.1006/anbe.1998.1049

684 Jennings, D. J., Carlin, C. M., & Gammell, M. P. (2009). A winner effect supports

685 third-party intervention behaviour during fallow deer, Dama dama, fights. *Animal*
686 *Behaviour*, *77*(2), 343–348. http://doi.org/10.1016/j.anbehav.2008.10.006

687 Kaburu, S. S. K., & Newton-Fisher, N. E. (2015). Egalitarian despots: hierarchy
688 steepness, reciprocity and the grooming-trade model in wild chimpanzees, Pan
689 troglodytes. *Animal Behaviour*, *99*, 61–71.
690 http://doi.org/10.1016/j.anbehav.2014.10.018

691 Kelstrup, H. C., Hartfelder, K., & Wossler, T. C. (2015). Polistes smithii vs. Polistes
692 dominula: the contrasting endocrinology and epicuticular signaling of sympatric
693 paper wasps in the field. *Behavioral Ecology and Sociobiology*, *69*(12), 2043–
694 2058. http://doi.org/10.1007/s00265-015-2015-9

695 Lusseau, D., Whitehead, H., & Gero, S. (2008). Incorporating uncertainty into the
696 study of animal social networks. *Animal Behaviour*, *75*(5), 1809–1815.
697 http://doi.org/10.1016/j.anbehav.2007.10.029

698 Majolo, B., Aureli, F., & Schino, G. (2012). Meta-analysis and animal social
699 behaviour. *Evolutionary Ecology*, *26*(5), 1197–1211.
700 http://doi.org/10.1007/s10682-012-9559-1

701 Martin, P., & Bateson, P. (2007). *Measuring Behaviour. An introductory guide.* (3rd
702 ed.). Cambridge: Cambridge University Press.

703 McDonald, D. B., & Shizuka, D. (2013). Comparative transitive and temporal
704 orderliness in dominance networks. *Behavioral Ecology*, *24*(2), 511–520.
705 http://doi.org/10.1093/beheco/ars192

706 Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution:
707 feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*, 88.

http://doi.org/10.1186/s12915-015-0196-3

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935–956. http://doi.org/10.1111/j.1469-185X.2010.00141.x

Neumann, C., Duboscq, J., Dubuc, C., Ginting, A., Irwan, A. M., Agil, M., … Engelhardt, A. (2011). Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating. *Animal Behaviour*, *82*(4), 911–921. http://doi.org/10.1016/j.anbehav.2011.07.016

Neumann, C., & Kulik, L. (2014). EloRating: Animal Dominance Hierarchies by Elo Rating. Retrieved from https://cran.r-project.org/web/packages/EloRating/index.html

Parker, G. A. (1974). Assessment strategy and the evolution of fighting behaviour. *Journal of Theoretical Biology*, *47*(1), 223–243. http://doi.org/10.1016/0022-5193(74)90111-8

Poisbleau, M., Guillon, N., & Fritz, H. (2010). Preservation of winter social dominance status in Brent Geese Branta bernicla bernicla within and across winters. *Journal of Ornithology*, *151*(3), 737–744. http://doi.org/10.1007/s10336-009-0437-8

Polačik, M., & Reichard, M. (2009). Indirect fitness benefits are not related to male dominance in a killifish. *Behavioral Ecology and Sociobiology*, *63*(10), 1427–1435. http://doi.org/10.1007/s00265-009-0798-2

Rat, M., van Dijk, R. E., Covas, R., & Doutrelant, C. (2015). Dominance hierarchies and associated signalling in a cooperative passerine. *Behavioral Ecology and*

731 *Sociobiology*, *69*(3), 437–448. http://doi.org/10.1007/s00265-014-1856-y

732 Roberts, S.-J., & Cords, M. (2015). Life as a bachelor: quantifying the success of an
733 alternative reproductive tactic in male blue monkeys. *PeerJ*, *3*, e1043.
734 http://doi.org/10.7717/peerj.1043

735 Sánchez-Tójar, A., Schroeder, J., & Farine, D. R. (2017). Supplementary material for
736 "A practical guide for inferring reliable dominance hierarchies and estimating
737 their uncertainty." http://doi.org/10.17605/OSF.IO/9GYEK

738 Sasaki, T., Penick, C. A., Shaffer, Z., Haight, K. L., Pratt, S. C., & Liebig, J. (2016). A
739 Simple Behavioral Model Predicts the Emergence of Complex Animal
740 Hierarchies. *The American Naturalist*, *187*(6), 765–775.
741 http://doi.org/10.1086/686259

742 Schielzeth, H., Stoffel, M., & Nakagawa, S. (2016). rptR: Repeatability Estimation for
743 Gaussian and Non-Gaussian Data. R package version 0.9.1.9000. Retrieved
744 from https://cran.r-project.org/package=rptR

745 Schjelderupp-Ebbe, T. (1922). Beiträge zur Sozialpsychologie des Haushuhns.
746 *Zeitsch F Psychol*, *88*, 226–252.

747 Sheppard, J. K., Walenski, M., Wallace, M. P., Vargas Velazco, J. J., Porras, C., &
748 Swaisgood, R. R. (2013). Hierarchical dominance structure in reintroduced
749 California condors: correlates, consequences, and dynamics. *Behavioral*
750 *Ecology and Sociobiology*, *67*(8), 1227–1238. http://doi.org/10.1007/s00265-
751 013-1550-5

752 Snyder-Mackler, N., Kohn, J. N., Barreiro, L. B., Johnson, Z. P., Wilson, M. E., &
753 Tung, J. (2016). Social status drives social relationships in groups of unrelated

754 female rhesus macaques. *Animal Behaviour*, *111*, 307–317.
755 http://doi.org/10.1016/j.anbehav.2015.10.033

756 Stewart, E. C., & Greives, T. J. (2016). Short-term immune challenge does not
757 influence social dominance behaviour in top-ranked black-capped chickadees.
758 *Animal Behaviour*, *120*, 77–82. http://doi.org/10.1016/j.anbehav.2016.07.023

759 Strandburg-Peshkin, A., Farine, D. R., Couzin, I. D., & Crofoot, M. C. (2015). Shared
760 decision-making drives collective movement in wild baboons. *Science*,
761 *348*(6241), 1358–1361. http://doi.org/10.1126/science.aaa5099

762 R Core Team (2016). R: A language and environment for statistical computing.
763 Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-
764 project.org/

765 von Rueden, C., Gurven, M., & Kaplan, H. (2008). The multiple dimensions of male
766 social status in an Amazonian society. *Evolution and Human Behavior*, *29*(6),
767 402–415. http://doi.org/10.1016/j.evolhumbehav.2008.05.001

768 Whitehead, H. (2008). Describing and modeling social structure. In *Analyzing animal*
769 *societies. Quantitative methods for vertebrate social analysis* (pp. 143–240).
770 Chicago: University of Chicago Press.

771 Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed*
772 *effects models and extensions in ecology with R*. New York: Springer.
773 http://doi.org/10.1007/978-0-387-87458-6

774