Inference of sexual selection from pairwise contests

Thomas Yu and Andrew G. Clark

Departments of Molecular Biology & Genetics and Computational Biology, Cornell University,
Ithaca NY, USA

¹Corresponding author: ac347@cornell.edu

Running title: Modeling pairwise contests

Authors' contributions

AGC conceived the study, TY performed the coding to format the data and obtain parameter estimates, and both authors wrote and edited the paper.

Acknowledgements

This work was supported by NIH R01 HD059060 to AGC and M. F. Wolfner. We thank Martin Wells for insightful comments in developing these ideas and on the manuscript. Ana Marija Jakšić, Arvid Ågren, Ian Caldas, Jullien Flynn and Manisha Munasinghe provided helpful suggestions for the manuscript.

Data accessibility

All scripts developed in this paper will be available on GitHub upon publication.

Abstract

Sexual selection, whether mediated by male-male competition, female choice, or male choice, is often realized through a series of pairwise contests. Experimental evolutionists have long applied male-choice and female-choice experiments to quantify the relative mating abilities of genotypes, an important component of fitness. Here, we consider *Drosophila* mate-choice experiments and apply two mathematical models, the Bradley-Terry and the Éló models, which have been explicitly designed to quantify the merits of contestants and establish a ranking based on results from pairwise contests. Methods such as the Bradley-Terry model and Éló ratings also provide a way to predict the outcome of contests that have not yet taken place. Leave-one-out validation approaches allow us to assess the utility of these models to quantify relative fitnesses, how well the models fit the data, and how well they perform at prediction. After applying these methods to *Drosophila* mate-choice experiments, we interpret the results in terms of sexual selection, and discuss the implications of departures of the data from good fits to the models.

Keywords: sexual selection, Bradley-Terry, Éló, *Drosophila*, pairwise contests

Introduction

Sexual selection is a potent driver of evolutionary change and a powerful motivating force in animal behavior (Andersson 1994). First introduced in *The Origin of Species* and then developed in *The Descent of Man and Selection in Relation to Sex*, the theory of sexual selection explains how certain characteristics impact mate choice and in turn, how the application of those mate choices has driven the evolution of phenotypic traits (Darwin 1859, 1896). Both males (Darwin 1896) and females (Clutton-Brock 2009) engage in intrasexual competition for mates and reproductive opportunities. In general, the genetic variance for mating success appears in many organisms to exceed the genetic variance for other classical components of fitness, such as viability (Bundgaard and Christiansen 1972), especially under some environmental conditions (Yun et al. 2018). This, coupled with the challenges in measuring sexual selection, motivate the need to consider statistical paradigms that can provide the best estimates of mating success.

A variety of empirical approaches have been applied in an effort to quantify sexual selection. One of the most common experimental protocols is to present virgin individuals of one sex with a pair of individuals of the opposite sex and observe the first mating that occurs (Arbuthnott et al. 2017; Dechaume-Moncharmont et al. 2013; Herdman et al. 2004). Such pairwise contests are somewhat contrived and rarely match a natural situation, but in aggregate, many such pairwise trials might be able to produce reasonable inferences about the ranked mating abilities of the genotypes in the study. The estimates typically have high variance, but with adequate replication they can produce reliably repeatable results.

The models of Bradley and Terry (1952) and Éló (1978) are two of the many probability models that estimate player ratings and fit a model that also allows estimation of the probability that one player (or team) beats another (Cattelan 2012; Király and Qian 2017). Player ratings are

an estimate of “ability” of each individual, and are calculated based on outcomes of past pairwise contests. Bradley and Terry (1952) derives and develops a formal statistical model for inference of ranking when the data consist of pairwise contests, and the Éló ranking system was developed later and independently to provide a means for ranking the performance of chess players. The Bradley-Terry and Éló models can be used to test whether there is a simple ranking of players that display strict stochastic transitivity. The models can be used to make probabilistic predictions of outcomes of matches that have not yet occurred. So if A has beaten B two-thirds of the time, and B has beaten C two-thirds of the time, then not only do we expect A to beat C (simple transitivity), but in fact the model predicts that A will beat C with probability 0.8 (simply solve for $C/(A+C)$ when $A/(A+B) = 2/3$, and $B/(B+C) = 2/3$). The question of transitivity of behaviors arises frequently in population biology (Shafir 1994), and while it seems like a rational approach, there are situations both theoretical and empirical that produce nontransitive behaviors (McNamara et al. 2014). Application of the Bradley-Terry model is less common, but it too has appeared in the population biology literature. Shev et al. (2014) applied the Bradley-Terry model to obtain fits to the dominance hierarchy among rhesus macaques engaging in pairwise contests. Muller et al. (2015) obtained estimates of host preference of saproxylic beetles with the Bradley-Terry model, and they assessed the effect of covariates on preferences with a method of recursive partitioning.

We would like to answer whether modeling approaches like Bradley-Terry, which estimate intrinsic abilities or ranks of each genotype, provides a useful means for estimating fitness parameters when the experimental data are in the form of pairwise contests (mate-choice tests). Such a quantitative approach would have value in assessing how many pairwise tests are needed to obtain reliable inferences of relative mating abilities. Parameters of the model assign a

value or “ability” to each genotype and these in turn can be given confidence limits. These estimates are readily interpreted as relative mating abilities, and the differences in mating abilities can easily be used to infer allele frequency dynamics in a population (assuming mating ability is the only component of fitness that varies). After we apply the Bradley-Terry and Éló models to mate-choice data from Arbuthnott et al. (2017), we determine how well each model fits. Such tests of goodness-of-fit assess whether sexual selection occurs such that there is an intrinsic ranking of genotypes, as opposed to the outcome of a contest being dependent on specific interactions among the pairs of genotypes. Caveats to this approach are outlined in the Discussion.

Materials and Methods

Fly mate-choice data

Many mate-choice experiments have been published using *Drosophila*, but we needed a study that considered all pairwise tests among a collection of multiple lines. *Drosophila* mate-choice data was generously provided by Devin Arbuthnott from the paper Arbuthnott et al. (2017). That study was designed as a male-choice test and sought to determine whether there was transitivity to the choosing by males among pairs of female genotypes. Generally such experiments with *Drosophila* are done as female-choice tests (Andersson 1994), and part of the idea of this study was to consider whether male-choice is an important phenomenon in flies. We use the data from pairwise contests with Oregon-R males and females from eleven lines from the *Drosophila* Genetic Reference Panel (DGRP): 138, 313, 362, 399, 517, 555, 707, 714, 738, 774, and 786. Note that, because each line is a distinct inbred genotype, we use the terms line and genotype interchangeably. In each trial, two females from eleven DGRP lines were presented to

a male from an Oregon-R stock. Altogether, there were 55 distinct pairwise tests, and a total of 550 mating trials were conducted. The line number of the female fly that successfully mated within a 2 hour observational period was recorded. Among the 550 mating trials, 392 concluded with a successful mating.

Data analysis

We used R version 3.5.2, Eggshell Igloo, for the data analysis (R Core Team 2018). The R package readxl (Wickham & Bryan 2019) was used to input the data, and R packages BradleyTerry2 (Turner & Firth 2012) and Éló (Heinzen 2019) were used to fit Bradley-Terry and Éló models, respectively. Error propagation was calculated through the R package propagate (Spiess 2018) using first-order Taylor expansion.

Model equations

The basic modeling challenge is to estimate, based on the results of a series of pairwise contests among our 11 genetic lines of flies, a vector of scores, α_i , for $i = 1$ to 11. With these α_i estimates, the formula for the probability of line i being chosen for mating over line j is

$$\Pr(i \text{ chosen over } j) = \frac{\alpha_i}{\alpha_i + \alpha_j} \quad (1)$$

These α_i scores, which apply to both the Bradley-Terry and Éló models, could in principle be used as a metric for fitness, in this case reflecting only the relative fitnesses with respect to sexual selection. In the Bradley-Terry model,

$$\alpha_i = e^{\lambda_i} \quad (2a)$$

where λ_i is the “ability” parameter of line i (Turner & Firth 2012). Given the data matrix of observed counts of wins and losses for each pairwise contest, Bradley and Terry (1952) derived the maximum likelihood estimators for the λ_i ’s.

In what essentially amounts to a rediscovery of the Bradley-Terry model, the Élő rating system in chess, makes use of sequential pairwise chess match results and provides players with a global rating. In this Élő ranking system,

$$\alpha_i = 10^{R_i/400} \quad (2b)$$

where R_i is the “rating” of line i (Coulom 2007). The Élő system provides a convenient way to update these ratings with each successive contest.

Transitivity

One important assumption of the Bradley-Terry model is that the contests have the feature of transitivity. Transitivity, an indication of “rational” behavior (Arbuthnott et al. 2017), occurs when, if A beats B, and B beats C, then A beats C. Weak stochastic transitivity occurs when, given $\Pr(A \text{ beats } B) \geq 0.5$ and $\Pr(B \text{ beats } C) \geq 0.5$ then $\Pr(A \text{ beats } C) \geq 0.5$ (Houston 1991). Strong stochastic transitivity occurs when, given $\Pr(A \text{ beats } B) \geq 0.5$ and $\Pr(B \text{ beats } C) \geq 0.5$ then $\Pr(A \text{ beats } C) \geq \max[\Pr(A \text{ beats } B), \Pr(B \text{ beats } C)]$ (Houston 1991). Equations 2a and 2b hold true only if there is also strong stochastic transitivity. Note that an assessment of fit to the Bradley-Terry model implies not only transitivity, but a quantitative fit to the mating probabilities. If all that we knew was that the results are transitive, this would not by itself produce an ability to quantitatively predict outcomes of mate choices.

Tests of goodness-of-fit

Deviance statistics: The R packages generate a null deviance and a residual deviance statistic to measure how well the model fits. A smaller deviance value indicates a better fitting model. The null deviance is the deviance corresponding to the model containing no fitted parameters, while the residual deviance corresponds to the model containing all rating parameters of fly lines. The difference between the residual and null deviances approximates a χ^2 distribution with 10 degrees of freedom. We conduct a chi-square test to find P -values and determine significance (in this case, indicating that the model produces a significantly better fit to the data than does the null model).

Leave-one-out tests: We assessed the performance of the Bradley-Terry model by systematically excluding the results of pairwise contests between two randomly chosen lines from the data. We fit the Bradley-Terry model to the incomplete data and use it to predict the probability of the contest between the lines in the excluded contest. We then compare with a chi-square test the predicted probability with the actual counts of mating successes, giving us an idea of how well the model predicts the results of *Drosophila* mating assays. There are 55 pairs among the 11 lines, and so this yields 55 chi-square tests, each with one degree of freedom.

Transitivity tests: Let A, B, and C be three different lines. We find all possible trios and order them such that $A > B$ and $B > C$ (A engages in more matings than B, and B engages in more matings than C). After temporarily excluding the result of A vs. C from the data, we generate a Bradley-Terry model and use it to predict the probability that the male will choose the female of strain A over strain C. We then compare this to the actual percentage of trials in which A is chosen over C. We conduct a binomial test to generate a P -value for each of the 165 tests (the number of distinct trios of lines possible from a set of 11 lines). Note that these tests are not independent of one another.

We use a Q-Q plot of P -values to assess the goodness-of-fit of the model. P -values are obtained from the transitivity tests described above. We then compare obtained P -values from the binomial test with a uniform distribution of P -values. Using this graph, we can observe if and by how much the P -values differ from the expected null distribution.

Élő ratings

The Élő rating method is a system used to find a numerical rank-order of individuals in a group (Albers and de Vries 2001). Originally created to rank chess players, Élő ratings can be and are used in a variety of contexts. Here, we apply Élő ratings to eleven DGRP fly lines. The equation for calculating a line's rating is

$$R'_i = R_i + k(O_i - E_i) \quad (3)$$

where R_i is the rating of line i , R'_i is the update rating of line i , k is the k -factor, O_i is the result for line i ($O_i=1$ if line i mates, $O_i=0$ if line i does not mate), and E_i is the predicted probability that line i mates. E_i is calculated using Equations 1 and 2a for the Bradley-Terry model and using Equations 1 and 2b for the Élő model. The k -factor determines how much the rating changes in response to competition results, and is chosen somewhat arbitrarily based on the expected total number of contests. Here we choose $k = 24$.

Predicted allele-frequency dynamics

Sexual selection results in changes in population allele frequencies, and so to interpret the meaning of different models of sexual selection, it might be useful to consider how the models differ with respect to their predicted allele frequency dynamics. In this case, we wanted to know whether the allele frequency dynamics differed between a simple model where the fitnesses were inferred to be a property of each line, in a simple ranked order, compared to the model where

specific pairwise contests imposed altered frequency dynamics. We do this by simulating genotype frequencies across multiple generations. Note that all lines in the experiments are inbred and are therefore homozygous. Thus, we ignore the complexities of the real genetics, and model them as if they are haploid clones competing.

In our first simulation, we use *Élő* abilities as a direct proxy for fitness, substituting into the haploid frequency recursion:

$$f'_i = \frac{f_i \alpha_i}{\sum_j f_j \alpha_j} \quad (4a)$$

where f_i is the frequency of genotype i in the current generation, f'_i is the frequency of genotype i in the next generation, and α_i is the *Élő* ability of genotype i . This is thus equivalent to a one-locus, multiple allele haploid selection model.

In the next simulation, we use Bradley-Terry abilities to compute the probabilities of line i being chosen for mating over line j . These abilities comprise a fitness matrix p_{ij} , specifying the probability that i wins in a contest between i and j . To calculate the frequencies over successive generations, we use the following recursion:

$$f'_i = \frac{\sum_j (f_i f_j p_{ij})}{\sum_i \sum_j (f_i f_j p_{ij})} \quad (4b)$$

where f_i is the frequency of genotype i in the current generation, f'_i is the frequency of genotype i in the next generation. This is still a haploid model, but the genotypes are considered in all possible pairs, and mating abilities are a property of mating pairs rather than of individual genotypes.

Results

Fitting Élo and Bradley-Terry models

We calculate the Élo ratings of the eleven DGRP lines by assigning each line an initial rating of 400 (as is customary in the common use of Élo ratings in international chess) and inputting each pairwise competition as a “match” between two lines. As the number of matches increases, the ratings should tend to stabilize, and plotting the ratings is one means to assess whether sufficiently many matches have been run to obtain stable results. Ratings were calculated using Equation 3 and the match order was randomized. Élo ratings of different lines diverge significantly, indicating a difference in “ability” between genotypes (Figure 1). However, the rank ordering of the ratings is still changing after nearly 400 trials. We conclude that the model fits would be improved by collecting even more mating trials, and when the ratings stabilize, we would arrive at the best estimates for the “ability” of each line.

Figure 1 - Élo ratings of lines calculated over successive mating trials or “matches”. The order of the pairwise tests was randomized, and after each successive male-choice test, the ratings of the winning and losing genotype are updated. Note that even after 392 mating trials, the rank ordering has still not settled down to stable ratings. The ending ratings do not depend on the preceding order of matings.

Next, we generate fits to the Bradley-Terry and Éló models using the data. Among the fitted parameters is the “ability” of each line, which is a measure of how well a line is expected to perform relative to other lines in the group. Higher abilities are directly related to a higher probability of winning the pairwise contest (see Equation 1). Thus, it seems logical to presume that these estimates of ability might be a useful quantitative measure of relative mating success. Estimates for the abilities of each line are found in Table 1. A key conclusion is that both models produce the same rank ordering of lines’ mating abilities (Figure S1). We are interested in exploring whether these are useful proxies for estimates of reproductive fitness.

Line	Éló Ability	Bradley-Terry Ability
517	0.22 (0.12)	0.035 (0.02)
313	0.27 (0.13)	0.043 (0.02)
707	0.37 (0.17)	0.058 (0.03)
138	0.92 (.38)	0.15 (0.06)
786	1.00	0.16 (0.07)
555	1.36 (0.54)	0.21 (0.08)
714	1.85 (0.73)	0.29 (0.12)
399	2.25 (0.85)	0.35 (0.14)
774	3.78 (1.55)	0.59 (0.23)
738	6.36 (2.71)	1.00
362	15.69 (7.49)	2.47 (1.06)

Table 1 - Abilities of fly lines according to the Éló and Bradley-Terry models, in rank order from lowest to highest. Both models generate parameters representing log-ability for all lines. Abilities are calculated using e^{λ_i} , where λ_i is the parameter of line i for the corresponding model. Standard errors are in parentheses. The log-abilities of line 786 in the Éló model and line 738 in the Bradley-Terry model are set

to 0 as a point of reference, and therefore do not have standard errors. Note that the rank order of the lines is identical under the two models.

Tests of goodness-of-fit

We report the null and residual deviances for the Bradley-Terry and Éló models in Table 2. The difference between the residual and null deviance is a measure of how well the complete model explains the data compared to the model that assumes that all genotypes have equal mating ability. The highly significant P-values for both the Bradley-Terry and Éló models indicates that the models allowing fly lines to have distinct mating success probabilities greatly improves the fit of the model to the data compared to the null model, which allows no differences among lines. This further implies that the models have improved predictive capability over the null. Table S1 shows the tests of significance of each predicted mating trial compared to the null hypothesis of equal mating probability.

	Bradley-Terry	Éló
Null Deviance	249.07 (55 df)	543.39 (391 df)
Residual Deviance	56.70 (45 df)	350.72 (381 df)
P-value	6.29×10^{-36}	5.44×10^{-36}

Table 2 - Null and residual deviances of Bradley-Terry and Éló models. The highly significant P-values implies that the models allowing the fly lines to have distinct mating abilities fits the data far better than the null model, which assumes no difference across the fly lines.

Next we conducted the transitivity tests described in the “Materials and Methods” section using all possible trios of lines. We compared the predicted probability of line A beating line C

(given A beats B and B beats C) with the actual percentage of A beating C in the data. Although many of the points have similar observed and expected values, some also display some deviation. We therefore performed binomial tests to assess the significance of these transitivity tests. We use a Q-Q plot to assess overall goodness-of-fit of these 165 tests by plotting the expected P -values against observed P -values taken from the binomial tests (Figure 2).

10 of the 165 (6.06%) tests were significant at $\alpha = 0.05$, a magnitude of deviation that is quite likely to occur by chance. We see from the plot that the points fit along the $y = x$ line with only minor deviation, indicating that the observed P -values conform reasonably well to the expected values. Another test of the predictions is to calculate the probabilities of outcomes in the case where all data except for one of the 55 contests is included, and contrast this to the observed mating success of that trial. Figures S2 and Table S2 show again that the model predictions fit the data well, and in fact only one of the 55 tests is rejected at $P < 0.05$.

Figure 2 - Q-Q plot from the leave-one-out trio tests. P -values are from binomial tests. The diagonal line represents the equation $y = x$.

Inferred allele-frequency dynamics

Simulations of the allele frequency dynamics under two different models of sexual selection illustrate how the parameterization of the model can make a difference. In the first model, we used the $\bar{E}l\bar{o}$ abilities as a proxy for the haploid fitness of each allele, applied Equation 4a, and plotted the allele frequency dynamics (Figure 3A). In this case, the line with the highest ability eventually becomes the sole remaining line. This is expected for one-locus multiple allele haploid selection. The initial frequency, here chosen at random, affects the number of generations it takes for this line to go to fixation, but the line with the highest ability always goes to fixation, regardless of starting conditions. The second model considered the case in which mating outcomes depend on all 55 pairwise outcomes. Here we used the predictions for the 55 pairwise contests as the fitness parameters in Equation 4b, and plotted the allele frequency dynamics for an arbitrary starting frequency. In this case, the line that ultimately wins depends on both the mating abilities and the initial frequencies. When initial frequencies are equal for all lines, the line with the highest ability outcompetes the others. But when initial frequency for the highest ability line is sufficiently low, another line may be the ultimate winner, as seen in Figure 3B.

Figure 3 - Allele frequency dynamics, considering the DGRP lines as haploid clones. A) Fitness is equated to the *Élő* abilities, and the line with the highest ability always goes to fixation regardless of starting conditions. B) If instead we consider the expected success rates from the full matrix of pairwise contests, the behavior is now frequency dependent, so if line 362 is not sufficiently common, it can be outcompeted. Initial frequencies were randomly generated.

Discussion

Mate choice tests have been widely used by evolutionary biologists to assess differences among genotypes in mating ability, an essential component of fitness. For the most part, inference of parameters of sexual selection from an aggregation of pairwise tests has been done

by simple heuristics, without any formal statistical models. Here, we have shown that models, such as the Bradley-Terry model can provide a rigorous way to infer ratings for each genotype that serve as excellent proxies for fitness estimates. These models do make specific assumptions about the data and the situation to which they are applied (Shev et al. 2014). First, there is assumed to be a true ranking of the participants, and that there is a transitivity in the winners of contests. Contests are assumed to be independent of each other, and there is assumed to be no context dependence to the contest outcomes. Given the excellent fit that *Drosophila* mate choice data show to the Bradley-Terry model, these assumptions appear to be satisfied, at least in the constrained laboratory environment.

There are several advantages to the application of the Bradley-Terry model to pairwise contest data in the context of evolutionary biology. A formal parametric model allows the assessment of the overall goodness-of-fit of the model to the data. This allows the identification of lines or assays that deviate from the model. Additionally, the quality of the data can be ascertained, particularly whether the experiment is sufficiently large and consistently executed, to determine whether stable and robust estimates of parameters have been achieved (Figure 1). Finally, the parameter estimates from the model explicitly present the researcher with the ability to predict the outcomes of other pairwise tests, allowing a powerful form of cross validation.

This study also raises some interesting questions related to the old “unit of selection” debate (Lewontin 1970). Sexual selection can be modeled with fitness parameters for each genotype, and the outcome of particular mating trial depends in some simple way on the parameters of each (*e.g.* Equation 1). In this case, the unit of selection is the single genotype, and the model has as many parameters as there are genotypes. Alternatively, the outcome of a mating trial could be modeled with a matrix of parameters whose entries represent all possible

pairs of genotypes, and the unit of selection is now the genotype-pair. This matrix of 55 parameters can also be calculated from the Bradley-Terry model. Interestingly the resulting allele frequency dynamics of these two models (Equations 4a and 4b) can be quite different, as is demonstrated in Figures 3a and 3b above. This presents the caveat that, despite the very good fit to the Bradley-Terry model, models of pairwise contests can still can yield different population-level inferences about allele dynamics. In addition, here we consider only the homozygous inbred lines, and of course in any experiment of allele dynamics in a population, heterozygotes would be formed, so many additional pairwise tests would have to be done to assess the mating ability of heterozygotes as well.

An important assumption of the Bradley-Terry model is that the parameters are invariant across time and other contexts. There are several reasons to suspect that strict application of the model to data like the mate choice experiments presented here will have limited value in predicting allele frequency dynamics in nature. First, the parameter estimates undoubtedly depend on many attributes of the environmental conditions, including diet, temperature, degree of crowding, etc.. A fly's true mating performance in the wild is not determined by its first choice test as a virgin, but must be integrated over many trials, providing an opportunity for learning and tuning of choices.

Mate choice tests among more than two genotypes provide the opportunity to test for transitivity of the results. Much has been made of the notion that transitivity in choice tests implies that the animals behaving this way are “rational.” This terminology stems from decision theory, and the logic is simple so long as the contrast to transitivity is randomness. But the problem with mate choice tests is that we initially have no idea what attributes are being distinguished to make the decisions, at least not until additional perturbation experiments are

done, such as knocking out sensory neurons, or altering cuticular hydrocarbons (Arbuthnott et al. 2017). But as soon as you realize that there are myriad traits that are being integrated to make mate choices, it is clear that totally rational decisions can produce nontransitivity, as soon as different attributes are applied in different contexts.

The Bradley-Terry model, and the many related models that fit large numbers of pairwise contests, are well poised have multiple fruitful applications in evolutionary biology. Joint analysis of male-choice and female-choice experiments would allow an assessment of the different degree to which matings are impacted by male behavior, female behavior, or “luck” (Gilbert and Wells 2019). At present, there is a general belief that female choice is the prominent factor in *Drosophila* sexual selection, but recent studies have shown that male choice is robustly repeatable (Arbuthnott et al. 2017; Edward and Chapman 2012). Finally, we note that the Bradley-Terry model is over 65 years old, and that inference and prediction in the context of pairwise tests is an area of active development, including Bayesian methods (Gilbert and Wells 2019) and machine learning approaches (Menke and Martinez 2008). Some of these methods even accommodate multi-dimensional attributes of the competing individuals, making them even more useful in a broad collection of evolutionary applications.

References

- Albers, P. C. H., and H. de Vries. 2001. Éló-rating as a tool in the sequential estimation of dominance strengths. *Animal Behaviour* 61:489–495.
- Andersson, M. B. 1994. *Sexual Selection*. Princeton University Press, Princeton, N.J.
- Arbuthnott, D., T. Y. Fedina, S. D. Pletcher, and D. E. L. Promislow. 2017. Mate choice in fruit flies is rational and adaptive. *Nat Comm.* 8:13953.
- Bradley, R.A. and M.E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons". *Biometrika*. 39 (3/4): 324.
- Bundgaard J., and F.B. Christiansen. 1972. Dynamics of polymorphisms. I. Selection components in an experimental population of *Drosophila melanogaster*. *Genetics* 71:439-60.
- Cattelan, M. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Stat. Sci.* 27: 412-433.
- Clutton-Brock, T. 2009. Sexual selection in females. *Animal Behaviour* 77:3–11.
- Coulom, R. 2007. Computing “Éló Ratings” of move patterns in the game of Go. *ICG* 30:198–208.
- Darwin, C. 1859. *On The Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life*. London :John Murray.
- Darwin, C. 1896. *The Descent of Man and Selection in Relation to Sex*. D. Appleton.
- Dechaume-Moncharmont, F.-X., M. Freychet, S. Motreuil, and F. Cézilly. 2013. Female mate choice in convict cichlids is transitive and consistent with a self-referent directional preference. *Front. Zool.* 10:69.

- Edward, D. A., and T. Chapman. 2012. Measuring the fitness benefits of male mate choice in *Drosophila melanogaster*. *Evolution* 66:2646–2653.
- Élő, A. 1978. *The Rating of Chessplayers, Past and Present*. Arco. ISBN 978-0-668-04721-0.
- Gilbert, D.E. and M.T. Wells. 2019. Ludometrics: Luck and how to measure it. *J. Quant. Anal. Sports* 15(3): 225–237.
- Heinzen, E. 2019. elo: Élő Ratings. <https://cran.r-project.org/web/packages/elo/elo.pdf>
- Herdman, E. J. E., C. D. Kelly, and J.-G. J. Godin. 2004. Male mate choice in the guppy (*Poecilia reticulata*): Do males prefer larger females as mates? *Ethology* 110:97–111.
- Houston, A. 1991. Violations of stochastic transitivity on concurrent chains: Implications for theories of choice. *J Exp Anal Behav* 55:323–335.
- Király, F. J., and Z. Qian. 2017. Modelling competitive sports: Bradley-Terry-Élő models for supervised and on-line learning of paired competition outcomes. arXiv:1701.08055.
- Lewontin, R.C. 1970. The units of selection. *Ann. Rev. Ecol. Syst.* 1:1-18.
- McNamara, J. M., P. C. Trimmer, and A. I. Houston. 2014. Natural selection can favour “irrational” behaviour. *Biology Letters* 10:20130935–20130935.
- Menke J.E. and T.R. Martinez. 2008. A Bradley–Terry artificial neural network model for individual ratings in group competitions. *Neural Comput & Applic* 17:175–186.
- Muller, J., B. Wende, C. Strob, M. Eugster, I. Gallenberger, A. Floren, I. Steffan-Dewenter, K. E. Linsenmair, W. W. Weisser and M. M. Gossner. 2015. Forest management and regional tree composition drive the host preference of saproxylic beetle communities. *Jour. Applied Ecol.* 52:753-762.
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Shafir, S. 1994. Intransitivity of preferences in honey bees: support for “comparative” evaluation of foraging options. *Animal Behaviour* 48:55–67.
- Shev A, K. Fujii, F. Hsieh, B. McCowan. 2014. Systemic testing on Bradley-Terry model against nonlinear ranking hierarchy. *PLoS ONE* 9(12): e115367.
- Spiess, A.-N. 2018. Propagate, a comprehensive R package to infer error propagation. <https://cran.r-project.org/web/packages/propagate/propagate.pdf>
- Turner, H., and D. Firth. 2012. Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software* 48:1–21.
- Wickham, H., and J. Bryan. 2019. readxl: R-package to read Excel Files. <https://cran.r-project.org/web/packages/readxl/readxl.pdf>
- Yun L, Chen PJ, Kwok KE, Angell CS, Rundle HD, Agrawal AF. 2018. Competition for mates and the improvement of nonsexual fitness. *Proc Natl Acad Sci U S A*. 115:6762-6767.

Supporting Information

Figure S1 - Comparison of expected probabilities for mate choices under both the Bradley-Terry and Élo model. Expected probabilities are calculated from Equation 1, where the α_{is} represent Bradley-Terry “ability” and Élo “ability,” respectively (see Table 1). Each point represents the expected probability of one of the 55 pairwise contests.

Figure S2 – Leave-one-out tests showing the stability of the Bradley-Terry estimates in predicting the outcome of each of the 55 pairwise matings when data from each specific pairwise test is removed. The x-axis indicates the observed proportion of time genotype i successfully mates for each of the 55 contests, and the y-axis is the predicted probability for the same (i, j) pair based on the parameter estimates with the data from the (i, j) pairwise mating tests removed. There were 55 pairwise contests, and so the plot has 55 points. The sample was relatively modest, so there is evident scatter, but only one of these 55 tests significantly rejects a goodness-of-fit binomial test (Table S2).

Table S1 - Matrix of expected probabilities from the Bradley-Terry model. P_{ij} is the probability that line i will be selected over line j . Probability ± 1 standard error. z-test for $H_0: P_{ij} = 0.5$ (ns: $P > 0.05$, *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$) Significance codes appear in parentheses.

	738	399	313	555	138	774	517	786	714	362	707
738	1.00	0.74 $\pm$ 0.07 (***)	0.96 $\pm$ 0.02 (***)	0.82 $\pm$ 0.06 (***)	0.87 $\pm$ 0.05 (***)	0.63 $\pm$ 0.09 (ns)	0.97 $\pm$ 0.02 (***)	0.86 $\pm$ 0.05 (***)	0.77 $\pm$ 0.07 (***)	0.29 $\pm$ 0.09 (*)	0.95 $\pm$ 0.02 (***)
399	0.26 $\pm$ 0.07 (***)	1.00	0.89 $\pm$ 0.06 (***)	0.62 $\pm$ 0.13 (ns)	0.71 $\pm$ 0.11 (ns)	0.37 $\pm$ 0.12 (ns)	0.91 $\pm$ 0.05 (***)	0.69 $\pm$ 0.12 (ns)	0.55 $\pm$ 0.13 (ns)	0.13 $\pm$ 0.06 (***)	0.86 $\pm$ 0.07 (***)
313	0.04 $\pm$ 0.02 (***)	0.11 $\pm$ 0.06 (***)	1.00	0.17 $\pm$ 0.09 (***)	0.23 $\pm$ 0.11 (*)	0.07 $\pm$ 0.04 (***)	0.56 $\pm$ 0.18 (ns)	0.21 $\pm$ 0.11 (**)	0.13 $\pm$ 0.07 (***)	0.02 $\pm$ 0.01 (***)	0.43 $\pm$ 0.17 (ns)
555	0.18 $\pm$ 0.06 (***)	0.38 $\pm$ 0.13 (ns)	0.83 $\pm$ 0.09 (***)	1.00	0.59 $\pm$ 0.14 (ns)	0.26 $\pm$ 0.11 (*)	0.86 $\pm$ 0.08 (***)	0.57 $\pm$ 0.14 (ns)	0.42 $\pm$ 0.14 (ns)	0.08 $\pm$ 0.04 (***)	0.79 $\pm$ 0.1 (**)
138	0.13 $\pm$ 0.05 (***)	0.29 $\pm$ 0.11 (ns)	0.77 $\pm$ 0.11 (*)	0.41 $\pm$ 0.14 (ns)	1.00	0.2 $\pm$ 0 0.09 (***)	0.81 $\pm$ 0.11 (**)	0.48 $\pm$ 0.15 (ns)	0.33 $\pm$ 0.13 (ns)	0.06 $\pm$ 0.03 (***)	0.72 $\pm$ 0.13 (ns)
774	0.37 $\pm$ 0.09 (ns)	0.63 $\pm$ 0.12 (ns)	0.93 $\pm$ 0.04 (***)	0.74 $\pm$ 0.11 (*)	0.8 $\pm$ 0 0.09 (***)	1.00	0.95 $\pm$ 0.03 (***)	0.79 $\pm$ 0.1 (**)	0.67 $\pm$ 0.12 (ns)	0.19 $\pm$ 0.09 (***)	0.91 $\pm$ 0.05 (***)
517	0.03 $\pm$ 0.02 (***)	0.09 $\pm$ 0.05 (***)	0.44 $\pm$ 0.18 (ns)	0.14 $\pm$ 0.08 (***)	0.19 $\pm$ 0.11 (**)	0.05 $\pm$ 0.03 (***)	1.00	0.18 $\pm$ 0.1 (**)	0.11 $\pm$ 0.06 (***)	0.01 $\pm$ 0.01 (***)	0.37 $\pm$ 0.17 (ns)
786	0.14 $\pm$ 0.05 (***)	0.31 $\pm$ 0.12 (ns)	0.79 $\pm$ 0.11 (**)	0.43 $\pm$ 0.14 (ns)	0.52 $\pm$ 0.15 (ns)	0.21 $\pm$ 0.1 (**)	0.82 $\pm$ 0.1 (**)	1.00	0.35 $\pm$ 0.13 (ns)	0.06 $\pm$ 0.03 (***)	0.73 $\pm$ 0.12 (ns)
714	0.23 $\pm$ 0.07 (***)	0.45 $\pm$ 0.13 (ns)	0.87 $\pm$ 0.07 (***)	0.58 $\pm$ 0.14 (ns)	0.67 $\pm$ 0.13 (ns)	0.33 $\pm$ 0.12 (ns)	0.89 $\pm$ 0.06 (***)	0.65 $\pm$ 0.13 (ns)	1.00	0.11 $\pm$ 0.06 (***)	0.84 $\pm$ 0.09 (***)
362	0.71 $\pm$ 0.09 (*)	0.87 $\pm$ 0.06 (***)	0.98 $\pm$ 0.01 (***)	0.92 $\pm$ 0.04 (***)	0.94 $\pm$ 0.03 (***)	0.81 $\pm$ 0.09 (***)	0.99 $\pm$ 0.01 (***)	0.94 $\pm$ 0.03 (***)	0.89 $\pm$ 0.06 (***)	1.00	0.98 $\pm$ 0.01 (***)

Table S2 - Matrix of mating success percentages from the data, along with tests of significance of goodness-of-fit of the leave-one-out tests. D_{ij} is the observed percentage of times line i was selected over line j in the pairwise mating trails, and P_{ij} is the model prediction for the same quantity. Binomial test for $H_0: D_{ij}=P_{ij}$ (ns: $p>0.05$, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$). Significance codes appear in parentheses.

	738	399	313	555	138	774	517	786	714	362	707
738	NA	0.9 (ns)	0.88 (ns)	0.78 (ns)	0.88 (ns)	0.56 (ns)	1 (ns)	0.83 (ns)	0.71 (ns)	0.4 (ns)	0.83 (ns)
399	0.1 (ns)	NA	0.88 (ns)	0.56 (ns)	0.75 (ns)	0.33 (ns)	1 (ns)	0.89 (ns)	0.67 (ns)	0 (ns)	0.83 (ns)
313	0.13 (ns)	0.13 (ns)	NA	0.14 (ns)	0.25 (ns)	0 (ns)	0 (ns)	0 (ns)	0.14 (ns)	0.1 (ns)	0.67 (ns)
555	0.22 (ns)	0.44 (ns)	0.86 (ns)	NA	0.29 (ns)	0.38 (ns)	1 (ns)	0.5 (ns)	0.43 (ns)	0 (ns)	0.86 (ns)
138	0.13 (ns)	0.25 (ns)	0.75 (ns)	0.71 (ns)	NA	0 (ns)	1 (ns)	0.33 (ns)	0.17 (ns)	0 (ns)	0.86 (ns)
774	0.44 (ns)	0.67 (ns)	1 (ns)	0.63 (ns)	1 (ns)	NA	0.5 (*)	0.83 (ns)	0.43 (ns)	0.22 (ns)	1 (ns)
517	0 (ns)	0 (ns)	1 (ns)	0 (ns)	0 (ns)	0.5 (*)	NA	0.4 (ns)	0.2 (ns)	0 (ns)	0 (ns)
786	0.17 (ns)	0.11 (ns)	1 (ns)	0.5 (ns)	0.67 (ns)	0.17 (ns)	0.6 (ns)	NA	0.43 (ns)	0 (ns)	1 (ns)
714	0.29 (ns)	0.33 (ns)	0.86 (ns)	0.57 (ns)	0.83 (ns)	0.57 (ns)	0.8 (ns)	0.57 (ns)	NA	0.2 (ns)	0.57 (ns)
362	0.6 (ns)	1 (ns)	0.9 (ns)	1 (ns)	1 (ns)	0.78 (ns)	1 (ns)	1 (ns)	0.8 (ns)	NA	1 (ns)
707	0.17 (ns)	0.17 (ns)	0.33 (ns)	0.14 (ns)	0.14 (ns)	0 (ns)	1 (ns)	0 (ns)	0.43 (ns)	0 (ns)	NA