Behavioral Ecology Advance Access published June 25, 2015

Behavioral Ecology

The official journal of the ISBE International Society for Behavioral Ecology

Behavioral Ecology (2015), 00(00), 1–6. doi:10.1093/beheco/arv091

Invited Ideas

Item Response Trees: a recommended method for analyzing categorical data in behavioral studies

Andrés López-Sepulcre,[a,b] Sebastiano De Bona,[b] Janne K. Valkonen,[b] Kate D.L. Umbers,[c] and Johanna Mappes[b]

[a]CNRS UMR 7618, Institute of Ecology and Environmental Sciences Paris (iEES), Université Pierre et Marie Curie, Case 237, 7 Quai St Bernard, 75005 Paris, France, [b]University of Jyväskylä, Centre of Excellence in Biological Interactions, Department of Biological and Environmental Science, Survontie 9 C, 40500 Jyväskylä, Finland, and [c]School of Science and Health, University of Western Sydney, Hawkesbury Richmond, NSW 2753, Australia

Received 23 October 2014; revised 26 May 2015; accepted 29 May 2015.

Behavioral data are notable for presenting challenges to their statistical analysis, often due to the difficulties in measuring behavior on a quantitative scale. Instead, a range of qualitative alternative responses is recorded. These can often be understood as the outcome of a sequence of binary decisions. For example, faced by a predator, an individual may decide to flee or stay. If it stays, it may decide to freeze or display a threat and if it displays a threat, it may choose from several alternative forms of display. Here we argue that instead of being analyzed using traditional nonparametric statistics or a series of separate analyses split by response categories, this kind of data can be more holistically analyzed using a generalized linear mixed model (GLMM) framework extended to binomial response trees. Originally devised for the social sciences to analyze questionnaires with multiple-choice answers, this approach can easily be applied to behavioral data using existing GLMM software. We illustrate its use with 2 representative examples: 1) repeatability in the measurement of antipredator display escalation and 2) the analysis of predator responses to prey appearance.

Key words: **behavioral analysis, categorical data, escalation, ethology, GLMM, item response theory, ordinal data, predator–prey interactions, R, repeatability, response trees.**

INTRODUCTION

Analyzing behavioral responses often poses statistical challenges because they rarely conform to the normality assumptions required for parametric tests (Lehner 1996; Martin and Bateson 2007). Traditionally, this problem has been circumvented with some success by implementing nonparametric statistics, which make few assumptions about the distribution of the data. Although nonparametric tests are widely used and have provided a convenient solution, it is widely recognized that they limit the interpretation of behavioral data in several ways. Standard nonparametric tests are often uninformative of the effect size, have low power to detect significant differences (Jennions and Møller 2003), and their algorithms often provide inaccurate estimates of P values for small samples (Mundry and Fischer 1998). Moreover, they allow little more than the simplest sampling designs, making it impossible to analyze hierarchical samples, complex correlation structures (e.g., genetic, spatial or temporal correlations), or repeated measures (e.g., to calculate repeatability).

Over the last few decades, generalized linear models and generalized linear mixed effect models (GLMMs) have emerged as powerful tools for data analysis in biology (Bolker et al. 2009). They provide great flexibility to model a variety of non-Gaussian responses common in behavioral research, including counts (e.g., number of matings), frequencies (e.g., pecks per minute), time spans (e.g., latency to escape), and dichotomous responses (e.g., escape or not) without the need to resort to nonparametric tests or transformations of difficult interpretation. The GLMM framework not only allows for flexibility in the type of the response variable but also in its covariance structure. This allows the incorporation of complex data structures such as hierarchical sampling (e.g., individuals within

Address correspondence to A. López-Sepulcre. E-mail: alopez@biologie.ens.fr.

families), repeated measures, and other types of nonindependent data (temporal and spatial correlations, genetic and phylogenetic relatedness, etc.).

Although the GLMM framework is increasingly used in behavioral ecology and ethology to account for many of the problems of quantitative data, a large amount of behavioral data is not strictly quantitative. Behavior is often recorded as a series of categories (e.g., from the species' ethogram), from which we wish to make quantitative statements. For example, one may record the intensity of the courtship display of a great-crested grebe (*Podiceps cristatus*) in terms of what stage in the stereotyped behavioral sequence it ended (approach, head shaking, weed-picking display, joint diving, or mating; Huxley 1914). While the descriptor is not quantitative in itself, this behavioral escalation has an underlying quantitative yet immeasurable driver (e.g., motivation). In other cases, recorded behaviors may reflect alternative tactics rather than degrees of escalation and as such do not show any ordered structure. For example, under threat of an intruder, a common lizard (*Lacerta vivipara*) may be submissive and either escape or freeze to avoid the intruder. Alternatively it may show dominance and either perform a typical push-up threat display or fight the intruder off (Gvozdík and Van Damme 2003). In the case of an escalation process, such as in the grebe example, we may traditionally use nonparametric tests on the ranking of the behaviors according to intensity (e.g., Rillich et al. 2011; Olofsson et al. 2012). In the case of alternative strategies, we may use contingency tables to evaluate associations between treatments and behavioral outcomes (Ruxton and Neuhäuser 2010). Yet these are limited to testing for an association and do not inform us of the size of effects. Moreover, they do not account for the covariance structure of the measured variables and limit the ability to analyze most realistic experimental designs, thus precluding the estimation of quantities such as intraindividual repeatability, spatiotemporal correlations, or heritabilities. Multinomial GLMMs do allow for the specification of complex correlation structures in the analysis of categorical outcomes and have recently been developed in the behavioral and evolutionary literature (Hadfield 2010; Hadfield and Nakagawa 2010; Dean et al. 2011). Multinomial GLMMs model the probability of occurrence of each of a set mutually exclusive behaviors, yet as explained above, it is often more insightful to interpret behavioral categories as the compounded result of a series of hierarchical decisions. For example, the probability of a common lizard showing a push-up display in the example above is the compound of 2 probabilities: the probability of displaying aggression times the probability of, given that, displaying the push-up rather than a direct threat. Those 2 probabilities bear a more straightforward biological meaning, and it may therefore be sensible to test hypotheses on those rather than the unconditioned multinomial probabilities of freeze, escape, push-up, and threat separately.

Psychologists and social scientists have worked on a solution to analyzing these types of data that behavioral ecologists often encounter. Item response theory (IRT; Rasch 1981) was originally designed to analyze responses to tests and questionnaires where the responses take the form of categorical multiple-choice answers. Originally, the interest lied in understanding the relationship between individual qualities and their performance in tests composed of a variety of questions or items. In its simplest version, the items could be scored dichotomously (e.g., correct–wrong), and a function was fit to relate the propensity to answer questions correctly to a given property of the individual tested. Further developments allowed the incorporation of polytomous outcomes (Ostini and Nering 2006). In its most recent formulations, item response models can be parameterized as GLMMs of binomial or multinomial data (De Boeck and Wilson 2004; De Boeck et al. 2011), where the relationship among the multinomial outcomes of different types of items (e.g., tasks, questions or scenarios) is explicitly modeled. We believe this type of modeling will prove extremely useful to scientists recording animal behavior. Recording the categories of behavior in an ethogram under different scenarios is analogous to filling in a questionnaire. For example, imagine the grebe in the above example is asked "Do you like that male?" and is given the choice to answer "Not interested," "Somewhat interested," "Very interested," corresponding to the display of different behaviors. The literature on IRT is extensive, covering a multitude of data types, and we very much encourage the readers to explore it (some good references include Embretson and Reise 2000; Baker 2001), but we here focus on one specific type of model that we find particularly promising in behavioral ecology: Item Response Trees (IRTrees; De Boeck and Partchev 2012).

IRTree models are designed to analyze multivariate responses that result from a decision tree of binary responses (De Boeck and Partchev 2012). Response trees conceptualize the outcome of a behavioral observation (or response) as the final product of a decision process. The decision process is conceptualized as a tree where each branching node represents a question, and the branches represent a binomial trial with 2 alternative responses. To illustrate, in the intruder lizard example a first node may ask whether to be submissive or aggressive and, conditional on that response, a second node asks whether to escape or freeze (in the case of submission), or an alternative third node asks whether to display or fight (in the case of having gone the aggressive route). The data is then analyzed as a multinomial response where the correlation structure is influenced by the relationship of potential outcomes in the node.

Conceptualizing the data as a correlated multinomial response allows it to be analyzed in a GLMM framework and thus take advantage of its flexibility in modeling a variety of experimental and sampling designs. The key to implementing binomial trees as GLMMs is to describe the response variable, decomposed in its component tree nodes, as a series of 0s or 1s depending on whether the response to any given node was positive or negative. That is, each observation will consist of as many binary data entries as there are nodes crossed in the tree path to the final behavioral outcome. The binomial entries should be grouped in the model by observation and node in order to account for their interdependency. The specification of random effects can allow the incorporation of a variety of experimental designs including repeated measures, time series, or hierarchical sampling, as well as correlation structures among nodes (e.g., whether individual responses to one node tend to correlate with their responses to another node).

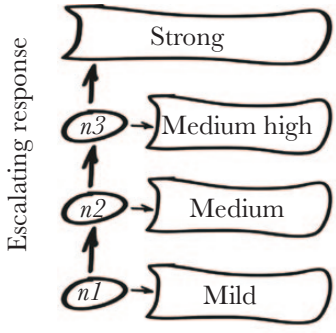
In this article, we use 2 examples to illustrate the use of IRTree GLMMs in behavioral research. The examples represent the 2 potentially common type of scenarios, as discussed above. First, we discuss an example of behavioral escalation in the mountain katydid *Acripeza reticulata*, where the categories can be ordinated. Second, we analyze an experiment on great tit *Parus major* responses to alternative prey, where the scored behaviors can be grouped into alternative paths. With the 2 examples, we will also exemplify how the random effects (e.g., individual variation) can be coded as affecting all nodes equally or differently. We will compare our proposed approach with more common nonparametric tests. For a detailed description on how to implement these analyses using standard GLMM packages in program R (R Core Team 2014), we have created an accompanying tutorial as Supplementary Appendix S1.

EXAMPLE 1: INDIVIDUAL REPEATABILITY OF KATYDID DEIMATIC DISPLAY

Mountain katydids (*A. reticulata*) inhabit Eastern Australia and are notable for their striking antipredator deimatic (i.e., startle) display (Umbers and Mappes 2015). Here we present data collected with the aim of measuring the repeatability of the antipredator response within individuals (Umbers and Mappes 2015) and the effect of desensitization on the response. In the experiment, 45 katydids were tapped twice consecutively on days 1, 3, and 5 of the experiment, and their behavioral response measured on a scale of 1–4, 1 being a mild response (antenna wiggling) and 4 being the highest level of escalation (full deimatic display where an individual holds its wings up and displays its vivid warning signals; Figure 1a).

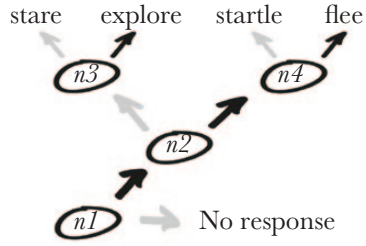
Traditionally, the repeatability of pairs of ordinal measurements is analyzed by calculating the Spearman rank correlation between the first and second measures of individuals on a given day (Martin and Bateson 2007). In our case, the Spearman's correlation is $\rho = 0.67$ ($P < 0.001$). The limitation of this approach is that, because it uses ranks, it lacks quantitative interpretation unless one assumes the different levels of the ordered category to be equidistant (e.g., the difference between "mild" and "medium" is the same as between "medium" and "medium high"). Moreover, it is not comparable with standard Anova-based repeatability measures (e.g., proportion of variance explained by individual) used for normally distributed variables. On the other hand, analogous measures can be calculated for non-Gaussian distributions if we can use a GLMM specification (Nakagawa and Schielzeth 2010).

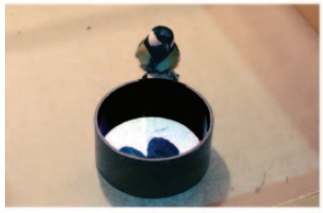
(a)



Random Effects	Variance			
Individual	1.54			
Observation	<0.01			
Fixed Effects	**Estimate**	**SE**	**Z-value**	**p-value**
Node 1 (Intercept)	2.65	0.32		
Node 2 (Intercept)	0.98	0.27		
Node 3 (Intercept)	-0.93	0.30		
Day	-0.34	0.08	-4.07	<0.01

(b)





Random Effects	Variance	Correlations		
Observation	< 0.01			
Individual		Node 1	Node 2	Node 3
Node 1	33.41			
Node 2	6.50	0.51		
Node 3	3.20	0.53	-0.46	
Node 4	12.21	-0.22	-0.95	0.71
Fixed Effects	**Estimate**	**SE**	**Z-value**	**p-value**
Node 1				
Intercept	3.61	1.64		
Eyespots	-0.32	1.80	-0.18	0.86
Node 2				
Intercept	-2.56	1.23		
Eyespots	3.39	1.39	2.43	0.02
Node 3				
Intercept	-0.84	1.03		
Eyespots	-2.80	1.85	-1.51	0.13
Node 4				
Intercept	1.03	2.05		
Eyespots	0.73	1.96	0.37	0.71

Figure 1
Diagrammatic representation of model structure and results from the 2 case studies. (a) Katydid deimatic response can escalate from mild to strong. This can be represented as a decision tree with 3 nodes where each node is associated with a binomial probability of escalating (1) or not (0). For example, a medium response corresponds to $n1 = 1$, $n2 = 0$, $n3$ and $n4$ remained undefined. The table shows the estimates of variance components and fixed effects. Note that the intercept is centered at day 1 so that they can be interpreted as the logit probabilities of escalating on the first day of observations. Pictures: 3 intensities of the katydid deimatic display; © Kate DL Umbers. (b) Great tit alternative responses to butterfly models are determined by a decision tree involving the probability of responding (node 1, $n1$), the probability of responding adversely (node 2, $n2$), the probability of exploring given that the response has been of interest (node 3, $n3$), and the probability of fleeing given an aversive response (node 4, $n4$). Black represents outcomes coded as 1 in the binomial trial, gray arrows represent outcomes coded as a 0. The tables show the model estimates, including node-specific individual random effects and their correlations, as well as treatment (eyespot) effects for all 4 nodes. Photo: great tit staring at a butterfly model; © Sebastiano De Bona. Note that for clarity, we only include z tests and P values for the effects of interest, and not the intercepts (for expanded results, see Supplementary Appendix S1).

To use the IRTree GLMM approach, the katydid's 4-level escalation can be envisioned as a sequence of binomial decisions with 3 nodes where the individual determines whether to go up one level of escalation or stay at their current level (Figure 1a). Because it is not known whether the defined behaviors represent a continuum on a linear scale, nodes can be assigned different baseline probabilities by including node identity as a fixed factor in the model. Such an approach allows for different intercepts for each node (i.e., different probabilities of escalating at different levels), therefore relaxing the assumption of equal distance between escalation levels. To test and account for the fact that individuals may desensitize with repeated stimulation, we include day of trial as a fixed covariate. Incorporating a node–day interaction could show whether the day effect is different as individuals go up the display intensity scale. However, for simplicity of illustration, we assume that day of observation has a similar effect on the escalation probabilities at all levels. Finally, because observations were performed twice a day and on different individuals, we include random effects for individual and replicate within each day (for full details, see Supplementary Appendix S1).

Including individual as a random effect quantifies the variance of the behavior (in this case, sequential probability of escalation) among individuals (see the table in Figure 1a). Repeatability can then be estimated as the proportion of variance explained by the individual (Nakagawa and Schielzeth 2010). Because the distribution-specific variance for a logit-binomial model is $\pi^2/3$ (Nakagawa and Schielzeth 2010), the individual repeatability of the escalating behavior is thus thus the individual variance divided by the total of all variance components plus the distribution's link specific variance $\frac{\sigma^2_{ind}}{\sigma^2_{total}} = \frac{\sigma^2_{ind}}{\sigma^2_{ind} + \sigma^2_{obs} + \frac{\pi^2}{3}} = 1.54/(1.54 + 0 + 3.28) = 0.42$ (Figure 1a), where σ^2_{ind} is the individual variance and σ^2_{obs} is the observation level variance (or additive overdispersion component sensu Nakagawa and Schielzeth 2010). The behavior is considerably repeatable because 42% of the variance can be explained by individual identity. The results also indicate a significant negative effect of trial day (-0.34 ± 0.08, $P < 0.01$; Figure 1a), suggesting that individuals may desensitize with the number of trials.

EXAMPLE 2: RESPONSES OF GREAT TITS TO BUTTERFLY EYESPOTS

Our second example, a subset of data from a larger experiment, aims at investigating the role of butterfly wing eyespots in deterring potential predators. In this experiment, individual great tits (*P. major*) were presented with animated computer images of the owl butterfly (*Caligo martia*) displaying (or not) eyespots on their wings (De Bona et al. 2015). Each bird was tested twice on a single treatment. The bird's behavior was recorded as either 1) nonresponsive, 2) approach, 3) exploration, 4) startle, or 5) flee.

Traditionally, nonordinated behavioral categories have been often analyzed using contingency tables (chi square or Fisher's Exact test), to test whether the association between treatments and behaviors significantly departs from random. For our data, both a chi-square approximation ($\chi^2 = 56.2$, $P < 0.001$) and Fisher's Exact test ($P < 0.001$) show strong evidence for a nonrandom association. However, this reveals us little about which behaviors were different and how, nor allows us to account for repeated measures of individuals.

To be analyzed as an IRTree GLMM, these 5 behavioral categories can be conceptualized as a nonordered sequential decision tree (Figure 1b) determined by 4 binomial nodes: 1) the probability of showing a response (node 1), 2) the probability of responding aversively (behaviors 4 and 5), rather than to show interest (behaviors 2 and 3, node 2), 3) whether to explore conditional on having shown interest (node 3), and 4) whether to flee given that the response was aversive (node 4).

In this case, we are interested in the effect of the treatment (butterfly eyespots), on each of the nodes separately, because they refer to qualitatively different processes. Thus, we include treatment and node as interacting fixed factors. For clarity, we set separate intercepts for each node. As individual birds were tested more than once, we need to include a random effect for individual, which can be node specific, because it is possible that individual variation is expressed differently at all nodes (for full details, see Supplementary Appendix S1).

The results are shown in the Figure 1b and reveal that the presence of eyespots on a butterfly wing only has an effect on node 2 (interest vs. aversion). In particular, the presence of eyespots increases the probability of showing an aversive response. The strength of the aversive response (node 4: flee vs. startle) does not seem to be affected. We can also assess from the random effect correlation structure (node by individual) the individual correlations. For example, the strong positive correlation (0.71) between nodes 3 and 4 indicates that individuals who tend to respond strongly (flee) when aversive, tend to also respond strongly (explore) when interested. In other words, some individuals tend to be more active than others, regardless of the type of behavior that the object elicits.

DISCUSSION

Behavioral data can often be conceptualized as a decision tree leading to alternative categorical outcomes. We believe this applies to a large range of phenomena that behavioral ecologists are interested in, including mate choice, social interactions, or antipredatory responses that are not easily analyzed using traditional approaches. In Table 1, we propose a list of hypothetical examples that could be analyzed using IRTrees. We have shown how, by conceptualizing the behavioral responses as decision tree, we can analyze such data using a GLMM framework. This provides a variety of advantages. First, by requiring to organize the recorded responses into biologically meaningful decision structures, it stimulates the researcher to decompose behaviors into their constituent components. Second, it allows for simultaneously testing those parts, accounting for the structural dependencies caused by the trees, and therefore avoiding problems associated with performing multiple tests. Third, it allows for complex experimental designs, such as repeated measures, or hierarchical sampling designs. We have shown examples where individual (random) effects can affect all the nodes equally (katydid probability of deimatic display) or separately (alternative responses to butterfly eyespots). The GLMM framework allows for the incorporation of more complex random effect structures than the ones shown here, such as genetic relatedness (i.e., the animal model) or temporal autocorrelation (time series of behaviors). This flexibility permits the measurement of important quantities such as individual repeatability (Nakagawa and Schielzeth 2010), heritability (Wilson et al. 2009), phylogenetic signal (Hadfield and Nakagawa 2010), or correlated behavioral responses (e.g., for the study of personalities), which are not possible to calculate using nonparametric alternatives. Finally, the parametric nature of the analyses allows for easy estimation of effect sizes.

Naturally, an increase in model complexity comes at the cost of requiring higher sample sizes. This is particularly important to consider in behavioral studies, where sample sizes tend to be considerably lower than poll-based studies for which IRT was

Table 1

Examples of hypothetical behavioral applications for IRTree models

Behavior Species	Question	Categorical responses	Explanatory variables
Escalating			
Courtship *Blue footed booby*	Does male foot color affect female interest?	Approach Feet display Sky calling Bill contact Mating	Foot color intensity as fixed effect Female and male identity as random effects
Contest *Field crickets*	Does previous winning experience affect willingness to escalate?	Ignore Antenna fencing Engagement Grappling	Number of previous wins as fixed effect Individual as random effect
Anti-predator *Cuttlefish*	Does reproductive status affect the probability of escalation?	Crypsis Startle display Escape jet Ink release	Reproductive status as fixed factor Individual as random effect
Cognition *Corvids*	Is there a phylogenetic signal of cognitive ability?	Increasing levels of complexity in tests	Individual, species and phylogenetic relatedness as random effects
Alternative			
Territoriality *Anole lizards*	Does size affect the type of behavior shown under territorial challenge? Does temperature affect the intensity?	Submissive Freeze Flee Aggressive Dewlap display Chase away	Size and temperature as fixed effects interacting with node Focal individual and intruder as random effects
Social interaction *Mongolian gerbil*	What is the heritability of dominance behavior?	Submissive Avoid Groom Dominant Chase Fight	Relatedness matrix as node-specific random effect
Antipredator *Dice snake*	Does the type of threat affect the antipredator strategy? Is the strategy repeatable within individuals?	Defense Mimic viper Feign death Attack Coil Strike	Threat type as node-specific fixed factor Individual as random effect
Mate choice *Jumping spiders*	Does male leg-flicking frequency influence reaction of females?	Aversion Chase Kill Interest Allow approach Mate	Male flicking frequency as node-specific fixed effect Male size as covariate Male and female identity as random effects

originally designed. We thus highly recommend the use of simulations tailored to assess the power of any given GLMM specification and study design where sample size is concerned (Johnson et al. 2015). In the context of IRTs, the clinical psychology literature on patient-reported outcomes, motivated by stronger limitations in sample size than other applications, has a healthy tradition of providing simulation studies to evaluate power and sample size for a variety of models of differing complexity (e.g., Holman et al. 2003; Sébille et al. 2010; Hardouin et al. 2011; Blanchin et al. 2013; Guilleux et al. 2014). The general message is that, as the models get more complex and the expected effects are smaller, appropriate sample sizes required to find significance grow from dozens (as in our examples) to a few hundred. Similar conclusions are reached by studies on classical multinomial GLMMs (e.g., Jiang and Oleson 2011). Other important insights to be gained from simulations of GLMMs include 1) the estimation accuracy and bias of the random effects (e.g., individual variation) for different numbers of within and between-group observations (van de Pol 2012) and 2) the influence of the random effect structure on the confidence of the estimate and rate of false positives (Schielzeth and Forstmeier 2009). Johnson et al. (2015) provide a detailed tutorial on how to simulate binomial GLMMs with ecological example. In Supplementary Appendix S2, we illustrate how to perform such simulations specifically for IRTree GLMM models, using the katydid and great tit examples.

Psychology and sociology have recently seen important developments in methods that handle the difficulties of categorical data (Powers and Xie 2008). Many challenges in these fields are common to behavioral ecology and ethology and thus provide exciting new avenues for behavioral ecologists (see Nettle and Penke 2010; Carter et al. 2013 for similar arguments regarding the study of animal personalities). Individual response trees are a good example of how this exchange could be highly beneficial. We hope that our article encourages their application to behavioral data and inspires a better communication of statistical advances across disciplines.

SUPPLEMENTARY MATERIAL

Supplementary material can be found at http://www.beheco.oxfordjournals.org/

FUNDING

Funding was provided by the Academy of Finland (project 252411 for Centre of Excellence programs), the Hermon Slade Foundation (HSF14/3 to K.D.L.U. and J.M.), and the Generalitat Catalana (Grup de Calitat SGR 481 to A.L.S.).

We would like to thank D. Abondano and H. Nisu for their help gathering the great tit data and D. Noble and S.P. Gordon for invaluable comments.

Editor-in-Chief: Leigh Simmons

REFERENCES

Baker F. 2001. The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation. College Park (MD): University of Maryland.

Blanchin M, Hardouin JB, Guillemin F, Falissard B, Sébille V. 2013. Power and sample size determination for the group comparison of patient-reported outcomes in Rasch family models. PLoS ONE. 8:e57279.

De Boeck P, Bakker M, Zwitser R, Nivard M. 2011. The estimation of item response models with the lmer function from the lme4 package in R. J Stat Software. 39:1–18.

De Boeck P, Partchev I. 2012. IRTrees: tree-based item response models of the GLMM family. J Stat Software. 48:1–28.

De Boeck P, Wilson M, editors. 2004. Explanatory item response models: a generalized linear and nonlinear approach. New York: Springer.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 24:127–135.

De Bona S, Valkonen JK, López-Sepulcre A, Mappes J. 2015. Predator mimicry, not conspicuousness, explains the efficacy of butterfly eyespots. Proc R Soc B. 282:20150202.

Carter AJ, Feeney WE, Marshall HH, Cowlishaw G, Heinsohn R. 2013. Animal personality: what are behavioural ecologists measuring? Biol Rev. 88:465–475.

Dean R, Nakagawa S, Pizzari M. 2011. The risk and intensity of sperm ejection in female birds. Am Nat. 178:343–354.

Embretson SE, Reise S. 2000. Item response theory for psychologists. Mahwah (NJ): Erlbaum Publishers.

Guilleux A, Blanchin M, Hardouin JB, Sébille V. 2014. Power and sample size determination in the Rasch model: evaluation of the robustness of a numerical method to non-normality of the latent trait. PLoS ONE. 9:e83652.

Gvozdík L, Van Damme R. 2003. Evolutionary maintenance of sexual dimorphism in head size in the lizard *Zootoca vivipara*: a test of two hypotheses. J Zool. 259:7–13.

Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. J Stat Softw. 33:1–22.

Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol. 23:494–508.

Hardouin JB, Amri S, Feddag ML, Sébille V. 2011. Towards power and sample size calculations for the comparison of two groups of patients with item response theory models. Stat Med. 31:1277–1290.

Holman R, Cees AW, de Haan RJ. 2003. Power analysis in randomized clinical trials based on item response theory. Contr Clin Trials. 24:390–410.

Huxley JS. 1914. The courtship habits of the Great Crested Grebe (*Podiceps cristatus*); with an addition to the theory of sexual selection. Proc Zool Soc Lond. 84:491–562.

Jennions MD, Møller AP. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behav Ecol. 14:438–445.

Jiang D, Oleson JJ. 2011. Simulation study of power and sample size for repeated measures with multinomial outcomes: an application to sound direction identification experiments (SDIE). Stat Med. 30:2451–2466.

Johnson PCD, Barry SJE, Ferguson HM, Müller P. 2015. Power analysis for generalized linear mixed models in ecology and evolution. Methods Ecol Evol. 6:133–142.

Lehner P. 1996. Handbook of ethological methods. 2nd ed. Cambridge: Cambridge University Press.

Martin P, Bateson P. 2007. Measuring behaviour: an introductory guide. 3rd ed. Cambridge: Cambridge University Press.

Mundry R, Fischer J. 1998. Use of statistical programs for nonparametric tests of small samples often leads to incorrect P values: examples from Animal Behaviour. Anim Behav. 56:256–259.

Nakagawa S, Schielzeth H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev. 85:935–956.

Nettle D, Penke L. 2010. Personality: bridging the literatures from human psychology and behavioural ecology. Phil Trans Roy Soc Lond B Biol Sci. 365:4043–4050.

Olofsson M, Eriksson S, Jakobsson S, Wiklund C. 2012. Deimatic display in the European swallowtail butterfly as a secondary defence against attacks from great tits. PLoS ONE. 7:e47092.

Ostini R, Nering ML. 2006. Polytomous item response theory models. New York: SAGE Publications.

van de Pol M. 2012. Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. Methods Ecol Evol. 3:268–280.

Powers DA, Xie Y. 2008. Statistical methods for categorical data analysis. 2nd ed. Bingley (UK): Emerald.

Rasch G. 1981. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.

Rillich J, Schildberger K, Stevenson PA. 2011. Octopamine and occupancy: an aminergic mechanism for intruder–resident aggression in crickets. Proc R Soc B. 278:1873–1880.

Ruxton GD, Neuhäuser M. 2010. Good practice in testing for an association in contingency tables. Behav Ecol Sociobiol. 64:1505–1513.

Schielzeth H, Forstmeier W. 2009. Conclusions beyond support: overconfident estimates in mixed models. Behav Ecol. 20:416–420.

Sébille V, Hardouin JB, Le Néel T, Kubis G, Boyer F, Guillemin F, Falissard B. 2010. Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients—a simulation study. BMC Med Res Methodol. 10:24.

Umbers KDL, Mappes J. 2015. Post-attack deimatic display in the mountain katydid *Acripeza reticulata*. Anim Behav. 100:68–73.

Wilson AJ, Réale D, Clements MN, Morrissey MM, Postma E, Walling CA, Kruuk LEB, Nussey DH. 2009. An ecologist's guide to the animal model. J Anim Ecol. 79:13–36.