Ecology, 94(7), 2013, pp. 1479–1486
© 2013 by the Ecological Society of America

Multilevel statistical models and the analysis of experimental data

Jocelyn E. Behm,[1,2,3,5] Devin A. Edmonds,[2,3] Jason P. Harmon,[4] and Anthony R. Ives[2]

[1] *Animal Ecology, Department of Ecological Science, Vrije Universitiet, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands*
[2] *Department of Zoology, University of Wisconsin, 430 Lincoln Drive, Madison, Wisconsin 53706 USA*
[3] *Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, 88 Xuefu Road, Kunming 650223 People's Republic of China*
[4] *Department of Entomology, North Dakota State University, P.O. Box 6050, Fargo, North Dakota 58108 USA*

Abstract. Data sets from ecological experiments can be difficult to analyze, due to lack of independence of experimental units and complex variance structures. In addition, information of interest may lie in complicated contrasts among treatments, rather than direct output from statistical tests. Here, we present a statistical framework for analyzing data sets containing non-independent experimental units and differences in variance among treatments (heteroscedasticity) and apply this framework to experimental data on interspecific competition among three tadpole species. Our framework involves three steps: (1) use a multilevel regression model to calculate coefficients of treatment effects on response variables; (2) combine coefficients to quantify the strength of competition (the target information of our experiment); and (3) use parametric bootstrapping to calculate significance of competition strengths. We repeated this framework using three multilevel regression models to analyze data at the level of individual tadpoles, at the replicate level, and at the replicate level accounting for heteroscedasticity. Comparing results shows the need to correctly specify the statistical model, with the model that accurately accounts for heteroscedasticity leading to different conclusions from the other two models. This approach gives a single, comprehensive analysis of experimental data that can be used to extract informative biological parameters in a statistically rigorous way.

Key words: Fejervarya limnocharis; *heteroscedasticity;* Microhyla fissipes; *parametric bootstrapping;* Polypedates leucomystax; *pseudo-replication; Xishuangbanna.*

Introduction

Analyzing even seemingly simple experimental data sets can present statistical challenges that limit the interpretation of experimental results. A first challenge in data analysis is identifying the dependencies among data points for a given experimental design. Data from ecological experiments are often underlain by complex correlations generated by a lack of independence, such as repeated measures on the same experimental units. Failure to account for non-independence when testing hypotheses leads to pseudo-replication in which standard statistics may falsely inflate the significance of a result. Multilevel (or mixed) models (Gelman and Hill 2008) are recently gaining attention in ecology (Qian and Shen 2007, Bolker et al. 2009) and provide a way to explicitly account for complex correlations in experimental data.

A second challenge is ensuring that the assumptions of the appropriate statistical model for the data are met. Ecological experiments may be susceptible to differences in variance among treatments, or heteroscedasticity (Cleasby and Nakagawa 2011); this may be especially true for studies involving phenotypic variation among individuals that is influenced by environmental quality. Improperly accounting for heteroscedasticity can cause bias and inconsistency in estimates of standard errors and *P* values (Hayes and Cai 2007). The issue of heteroscedasticity is well recognized in classic linear regression (Sokal and Rohlf 1981). Nonetheless, although there are no technical hurdles for accounting for heteroscedasticity in multilevel models (Zuur et al. 2009), this is apparently rarely done (Cleasby and Nakagawa 2011), and standard statistical packages such as those in the statistical language R do not make this easy.

Once the appropriate statistical model is identified and the assumptions are met, a third challenge is extracting the appropriate information from the statistical results. Even when a model is statistically appropriate for a given experimental design, often the statistical results (such as estimates of treatment effects) are not the target of interest of the overall study. Key information may lie in complex comparisons among treatments that cannot be made using a simple ANOVA framework. A common approach in the literature is to analyze a subset of the experimental data and only include treatments of interest. It is conceptually more coherent, and potentially more statistically powerful, to

Manuscript received 19 November 2012; revised 20 March 2013; accepted 28 March 2013. Corresponding Editor: J. M. Levine.

[5] E-mail: jebehm@wisc.edu

analyze the entire data set simultaneously, using a model that extracts the contrasts of interest directly.

Experiments involving species interactions may be particularly subject to these issues of non-independence, heteroscedasticity, and complex contrasts. When treatments involve multiple species in the same experimental arena, their responses are not independent. Depending on the ecological processes being studied, individuals from the same species may have markedly different responses to each experimental treatment, which could result in heteroscedasticity. Finally, the goals of such experiments are not simply to conclude that the treatments differ, but to describe how they differ, necessitating contrasts.

Here, we provide a statistical framework for incorporating heteroscedasticity into multilevel models followed by parametric bootstrapping to calculate the statistical significance of complex contrasts among treatment effects. The strength of this framework is that the entire data set can be analyzed with a single statistical model. We apply this framework to a data set from an interspecific competition experiment on three tadpole species from tropical Xishuangbanna prefecture, Yunnan Province, China. These were the most common species recorded during ephemeral pool surveys (J. E. Behm, X. D. Yang, and J. Chen, *unpublished manuscript*), and our objective was to compare the relative strengths of interspecific and intraspecific competition. The issue of non-independence arises because in treatments containing two species, information used to calculate the competitive effect of one species on the other is taken from the same experimental unit as that used to measure interspecific competition in the opposite direction. We also had heteroscedasticity: one species had high variability in its response to high competition relative to the other treatments. Finally, we wanted to assess the relative magnitudes of inter- to intraspecific competition, which is not given by a simple comparison between treatments.

We describe our framework in detail in *Methods*. To illustrate the importance of selecting the correct model and incorporating heteroscedasticity, we substitute two other "wrong" models into our framework and contrast the results to the "correct" model. Multilevel models allow for the analysis of data at both the individual and experimental unit level. Our first model analyzes data at the individual level, the second analyzes data at the basin level, and the correct model analyzes data at the basin level while incorporating heteroscedasticity.

Methods

Experimental design.—Our experiment was conducted during the rainy season at the Xishuangbanna Tropical Botanical Garden (XTBG), Yunnan Province, China. The three tadpole species we used, *Fejervarya limnocharis*, *Microhyla fissipes*, and *Polypedates leucomystax* (hereafter referred to as F, M, and P), are by far the most abundant species in the landscape and co-occur in the same ephemeral pools (J. E. Behm, X. D. Yang, and J. Chen, *unpublished manuscript*). We used a reduced response surface experimental design (Inouye 2001) to create inter- and intraspecific competition treatments. Competition treatments consisted of each species grown alone at low density (60 individuals per 35 L basin), alone at high density (120 individuals), and all three pairwise combinations of species at high density (60 individuals per species, 120 per basin), for a total of nine treatments. These densities were within the range observed in the field, and the low-density treatment was chosen to reduce, but not eliminate, competition within the bounds we observed in the field. Each treatment was replicated in four blocks in an outdoor laboratory. For each block, we collected egg masses from four ephemeral pools and pooled clutches from each species in large basins for hatching. Once tadpoles began to feed (Gosner stage 25 [Gosner 1960]), they were counted and haphazardly assigned to treatments. Due to a low abundance of eggs, we could not start all four blocks simultaneously. Blocks 1 and 2 began on 2 July, block 3 began on 10 July, and block 4 began on 19 July 2009. Each basin received an initial inoculum of planktonic algae, ground spirulina-based fish flakes every 3 d, and partial water changes (15–25 L) four times throughout the experiment. At the end of the experiment on day 21, all remaining tadpoles were euthanized and preserved in 90% ethanol. Our four response variables were mass, length, developmental stage, and survival. These are standard response variables recorded in tadpole competition studies (e.g., Parris and Semlitsch 1998, Smith et al. 2004), as increased competition generally results in decreased access to food, which affects growth, development, and survival. If tadpoles do survive in a competitive environment, size at and timing of metamorphosis can have large effects on fitness (Altwegg and Reyer 2003).

Statistical analyses.—Our statistical framework consists of three parts: (1) perform multilevel regressions to obtain coefficients for the effects of species' identities and treatment on the response variable; (2) combine these coefficients to extract information about the relative strengths of intra- and interspecific competition, and competition coefficients; and (3) use parametric bootstrapping to conduct statistical tests of the competition strengths. In presenting our approach, we focus on the conceptual structure of the statistical models; annotated code in the R statistical computing language (version 2.13.0 [R Development Core Team 2011]) is available in the Supplement.

To investigate different ways of treating the data, we used three multilevel regression models: (1) at the scale of individuals, (2) at the scale of replicates while assuming homogeneous variances, and (3) at the scale of replicates accounting for heteroscedasticity. Comparing models 1 and 2 addresses the issue of whether analyses should be conducted at the individual vs. replicate level, and comparing models 2 and 3 addresses the issue of heteroscedasticity. Although the models are constructed as regressions, they are similar to ANOVAs because treatment and species are categorical variables.

The structure of all three models is:

$$
\begin{aligned}
&Y_i = \alpha_{\mathrm{sp}[i]} + \beta_{\mathrm{trt}[i]} + \alpha\beta_{\mathrm{sp}[i],\mathrm{trt}[i]} + c_{\mathrm{blk}[i]} + d_{\mathrm{sp}\,|\,\mathrm{rep}[i]} + e_i \\
&c_{\mathrm{blk}[i]} \sim \mathcal{N}(0, \sigma^2_{\mathrm{blk}}) \\
&d_{\mathrm{sp}\,|\,\mathrm{rep}[i]} \sim \mathcal{N}\left(0, \sum_{\mathrm{sp}\,|\,\mathrm{rep}}\right) \\
&e_i \sim \mathcal{N}(0, \sigma^2)
\end{aligned}
\qquad (1)
$$

where Y_i is the response variable that is standardized to have mean 0 and variance 1; this standardization facilitates comparison of regression coefficients for different response variables. For the individual-level model (model 1), Y_i ($i = 1, \ldots, N_i$) is known for all N_i individuals that survived to the end of the experiment, while for the two replicate-level models Y_i ($i = 1, \ldots, N_{\mathrm{b}}$) is the mean response variable for all N_{b} replicates (basin). The fixed effect $\alpha_{\mathrm{sp}[i]}$ is a categorical variable with three levels corresponding to the three species, with the function sp[i] mapping the datum i onto the identity of the species. The fixed effect $\beta_{\mathrm{trt}[i]}$ is defined to identify the four levels of treatment for each species: 60 individuals of the focal species (for which Y_i is measured) occurring with 60 F, 60 M, 60 P, or 0 other individuals. In this coding, the same treatment has a different meaning for the focal species; for example, if F is the focal species, the treatment with 60 F gives intraspecific competition, whereas if M is the focal species, the same treatment corresponds to interspecific competition. To account for differences in responses of species to treatments, the interaction term $\alpha\beta_{\mathrm{sp}[i],\mathrm{trt}[i]}$ is included. The random effect $c_{\mathrm{blk}[i]}$ represents variation among blocks that affect all individuals/species, with this variation assumed to follow a Gaussian distribution with mean 0 and variance σ^2_{blk}. The random effect e_i is the residual variance among data points.

The random effect $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ distinguishes the three models. For model 1, $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ accounts for covariances among individuals within the same basin. The model assumes that a value of $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ follows a Gaussian distribution for each species within each basin with species-specific variance σ^2_s ($s = 1, 2, 3$); thus, within a given basin, all individuals of the same species have relatively high or low expected values of the response variable as a result of the treatment, with additional variation among individuals incorporated into the residual random effect e_i. The species-specific variances σ^2_s are the diagonal elements of the 3 × 3 covariance matrix $\boldsymbol{\Sigma}_{\mathrm{sp}\,|\,\mathrm{rep}}$. For individuals from different species within the same basin, the values of $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ for two species (s and s') have covariances $\sigma^2_{s,s'}$, which are given by the off-diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{sp}\,|\,\mathrm{rep}}$. Thus, the values of $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ account for non-independence of conspecific individuals within the same basin; non-independence of heterospecific individuals is similarly given by the covariances in $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$.

For model 2, the random effect $d_{\mathrm{sp}\,|\,\mathrm{rep}[i]}$ and its 3 × 3 covariance matrix $\boldsymbol{\Sigma}_{\mathrm{sp}\,|\,\mathrm{rep}}$ are defined as in model 1, but now the variances and covariances refer to differences among the values of the mean of the response variable in each basin. Thus, the diagonal elements of $\boldsymbol{\Sigma}_{\mathrm{sp}\,|\,\mathrm{rep}}$ give the species-specific variances σ^2_s ($s = 1, 2, 3$), and the off-diagonal elements $\sigma^2_{s,s'}$ give the covariances in mean responses of two species s and s' among basins.

Model 3 is similar to model 2, although heterogeneity of variances is incorporated among the 12 species-treatments, instead of only allowing differences in the variances for the three species. Thus, the covariance matrix $\boldsymbol{\Sigma}_{\mathrm{sp}\,|\,\mathrm{rep}}$ is 12 × 12, with the diagonal elements giving the variances σ^2_{st} (st = 1, . . . , 12) corresponding to each species-treatment. The off-diagonal elements give the covariance $\sigma^2_{\mathrm{st},\mathrm{st}'}$ between species-treatments; these will all be zero except for the three interspecific treatments (F + M, F + P, M + P). Note that all three models are multilevel and take into account correlations among individuals/species within the same basin; none would be a priori inappropriate for this experimental design.

Once the coefficients have been estimated from any of the three regression models, they can be combined to estimate different attributes of competition. Suppose, for example, we were interested in the competitive effect of M on F relative to the intraspecific competitive effect of F on itself. To simplify notation, let α_{F} denote the value of $\alpha_{\mathrm{sp}[i]}$ for species F, and β_{F}, β_0, and β_{M} denote the value of $\beta_{\mathrm{trt}[i]}$ when F experiences, respectively, high intraspecific competition (high-density treatment), low intraspecific competition (low-density treatment), and competition from M. The intraspecific effect, which we denote C_{FF}, is given by

$$
\begin{aligned}
C_{\mathrm{FF}} &= [\alpha_{\mathrm{F}} + \beta_{\mathrm{F}} + \alpha\beta_{\mathrm{FF}}] - [\alpha_{\mathrm{F}} + \beta_0 + \alpha\beta_{\mathrm{F0}}] \\
&= \beta_{\mathrm{F}} + \alpha\beta_{\mathrm{FF}} - \beta_0 - \alpha\beta_{\mathrm{F0}}.
\end{aligned}
\qquad (2)
$$

Here, the first term in brackets gives the response variable of F at high density, and the second term in brackets is the response variable of F at low density, so the difference between them is the effect of intraspecific competition. Similarly, the interspecific effect of M on F is

$$
\begin{aligned}
C_{\mathrm{FM}} &= [\alpha_{\mathrm{F}} + \beta_{\mathrm{M}} + \alpha\beta_{\mathrm{FM}}] - [\alpha_{\mathrm{F}} + \beta_{\mathrm{F}} + \alpha\beta_{\mathrm{FF}}] \\
&= \beta_{\mathrm{M}} + \alpha\beta_{\mathrm{FM}} - \beta_{\mathrm{F}} - \alpha\beta_{\mathrm{FF}}
\end{aligned}
\qquad (3)
$$

which gives the difference in the response variable for F between treatments with and without M competitors. Similar computations give estimates for all nine combinations of intra- and interspecific competition.

To obtain estimates of the competition coefficient that gives the strength of interspecific competition of M onto

TABLE 1. Strengths of intra- and interspecific competition (C), and competition coefficients (A), estimated for log mass of three tadpole species (F, M, and P) in competition experiments.

	Model 1		Model 2		Model 3	
Parameter	Estimate	SE or CI	Estimate	SE or CI	Estimate	SE or CI
Strength of competition						
C_{FF}	1.07*	0.45	1.07*	0.48	1.07*	0.17
C_{FM}	0.57	0.46	0.57	0.48	0.57*	0.22
C_{FP}	0.78†	0.50	0.61†	0.47	0.61	0.65
C_{MM}	1.25*	0.42	1.25*	0.44	1.25*	0.42
C_{MF}	0.93*	0.42	0.93*	0.44	0.93*	0.30
C_{MP}	0.64†	0.42	0.64†	0.44	0.64*	0.30
C_{PP}	0.51*	0.23	0.51*	0.26	0.51†	0.36
C_{PF}	0.25	0.24	0.34†	0.26	0.34	0.41
C_{PM}	0.40*	0.23	0.42*	0.26	0.42	0.34
Competition coefficients						
A_{FM}	0.61	(0.25, 1.47)	0.61	(0.24, 1.55)	0.61	(0.36, 1.02)
A_{FP}	0.75	(0.30, 1.89)	0.63	(0.25, 1.58)	0.63	(0.17, 2.35)
A_{MF}	0.72	(0.32, 1.66)	0.72	(0.31, 1.77)	0.72	(0.29, 1.84)
A_{MP}	0.55	(0.23, 1.20)	0.54	(0.24, 1.32)	0.54	(0.22, 1.40)
A_{PF}	0.77	(0.48, 1.24)	0.84	(0.50, 1.42)	0.84	(0.42, 1.72)
A_{PM}	0.89	(0.56, 1.42)	0.91	(0.54, 1.55)	0.91	(0.52, 1.59)

Notes: Three regression models were used: model 1 using individual-level data; model 2 using replicate-level data assuming homogeneity of variances among treatments; and model 3 using replicate-level data accounting for heteroscedasticity. SE is reported for the strength of competition; CI is reported for competition coefficients.

* $P < 0.05$; † $P < 0.1$ from parametric bootstrapping.

F relative to intraspecific competition of F onto itself, A_{FM}, we use

$$A_{FM} = \exp(C_{FM} - C_{FF}) \\ = \exp\Big(\beta_M + \alpha\beta_{FM} + \beta_0 + \alpha\beta_{F0} - 2(\beta_F + \alpha\beta_{FF})\Big). \tag{4}$$

We use the exponent here so that when inter- and intraspecific competition have the same strength (i.e., $C_{FM} = C_{FF}$), $A_{FM} = 1$. Values of $A < 1$ imply interspecific competition is weaker than intraspecific competition.

Hypothesis tests for the values of C and A involve linear combinations of the regression coefficients, and in the case of A, a nonlinear (exponential) function. Although there are standard asymptotic approaches for joint hypothesis tests of regression coefficients such as the glht function in the multcomp R package (Hothorn et al. 2008), parametric bootstrapping (Efron and Tibshirani 1993) is more flexible and secure. Asymptotic approximations may be poor for even moderate sample sizes and especially if samples are not independent; for example, the P values we obtained using the glht function were incorrect when we applied it to data simulated to be similar to our experimental data set. Therefore, we used parametric bootstrapping. This involves first fitting a given model to the data, then simulating a large number of data sets from the fitted model, and finally estimating the coefficients from the simulated data sets to compute values of C and A. The resulting distributions of simulated estimates give the approximate distribution of the estimators of C and A from which confidence intervals and P values can be obtained. Although we opted to keep these analyses in the frequentist domain, it is also possible to perform comparable analyses in the Bayesian domain (Clark 2007, McCarthy 2007).

RESULTS

We analyzed the entire data set consisting of three species (F, M, and P) and four treatments (low density, high density, interspecific with the two other species) with four blocks in a single analysis. A total of 3618 tadpoles were used to initiate the experiment (not 3600, due to counting errors), but only 2544 tadpoles survived to the end. For simplicity, we use log mass as the response variable to illustrate the results of the three models below, although the statistical issues we describe were similar for the other three variables (length, stage, and survival).

Individual-level vs. replicate-level analyses.—We contrasted individual and replicate-level analyses using regression models 1 and 2 to calculate the strengths of intra- and interspecific competition, C, and the competition coefficient, A (Table 1). The estimated parameters differed slightly between models, yet both identified the same parameters as being statistically significant. Finding that the individual-level model had no more power than the replicate-level model might seem surprising, because model 1 treated all 2544 individuals as data points, whereas model 2 treated each of the 48 replicates as data points (36 basins total, with the 12 basins containing two species counted twice). Nonetheless, the explanation is simple. Model 1 (appropriately) accounts for the fact that the individuals in the same basin were

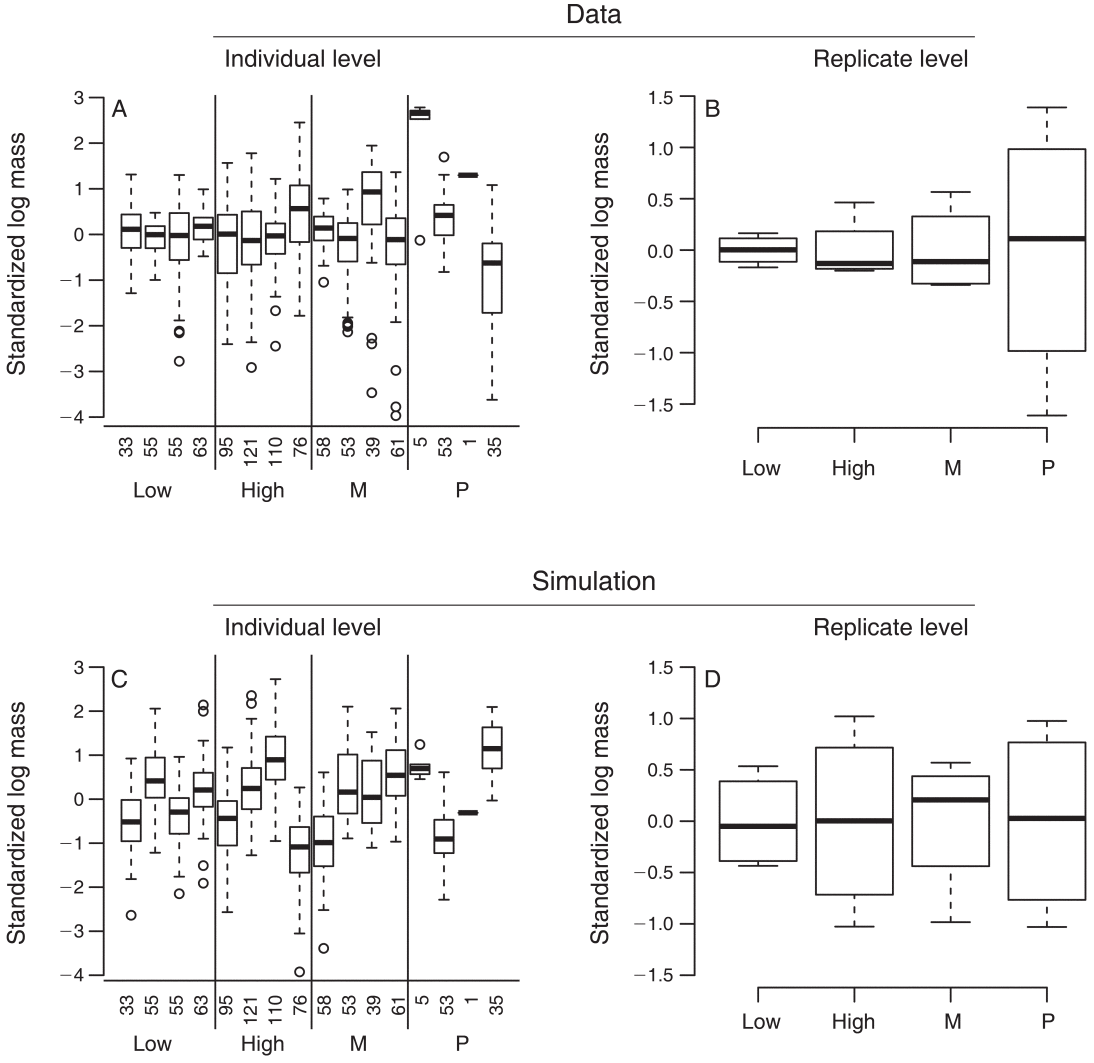

FIG. 1. (A, B) Variation in standardized log mass for species F from four experimental competition treatments (low and high density of conspecifics, and interspecific competition with M and P) for (A) the four replicates per treatment and (B) the mean across all four replicates within the treatment. (C, D) Similar representation of variation in log mass at the (C) individual replicate and (D) mean replicate levels for an example of simulated data generated using the individual model. Boxes in plots encompass the lower and upper quartiles of the data with the median represented by the thick horizontal line. Whiskers extend to the lowest and highest datum within 1.5 times the interquartile range. Numbers along the *x*-axis in plots A and C indicate the number of F individuals surviving in those replicates. Note that there is strong heteroscedasticity (variances are not equal) in the experimental data (panel B) and that this is greatly reduced (but not eliminated) in the simulated data (panel D).

all subjected to the same treatment. Because there is no way to find differences in responses to treatments among individuals within the same basin, in the statistical computations they show perfectly correlated responses, so more individuals add no new information to increase the statistical power.

What explains the slight differences between models in parameter estimates? After fitting model 1, we simulated data from the model under the assumption that all individuals survived within replicates. Applied to these simulated data, models 1 and 2 gave exactly the same results (not presented). For the real data, however, survival was low in some treatments; for example, two of the four replicates in the treatment for interspecific competition on F by P had one and five survivors. Low numbers of individuals per replicate will lead to greater measurement error. To assess the consequences of this, we simulated data using model 1 factoring in the number of survivors observed in the data (Fig. 1). These simulations show heteroscedasticity; treatment variances are unequal (Fig. 1D) even though all simulated individuals of a given species have the same variance in log mass. The difference in parameter estimates between models 1 and 2 is caused by differences in sample sizes (surviving individuals) among replicates.

Reports

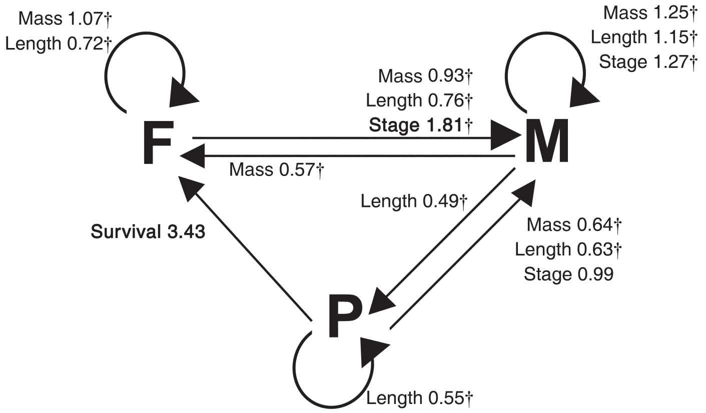

Fig. 2. Comparison of statistically significant competition strengths for the four response variables: mass, length, developmental stage, and survival from model 3. Arrows point from the species causing the effect to the species receiving the effect. All interactions that are significant for each variable are included, and those that remain significant after a Bonferroni correction for four comparisons (four response variables; $P < 0.0125$) are indicated with daggers (†). Response variables in boldface type also have significant competition coefficients for that interaction.

Heteroscedasticity.—Comparisons between the simulated data and the real data also showed that there is greater heteroscedasticity among treatments than can be explained by differences in sample sizes (Fig. 1). We performed 20 000 simulations of model 1 using the observed survival of individuals (illustrated by the example in Fig. 1C, D) and compared these to the observed data (Fig. 1A, B). The data showed greater heterogeneity among replicates in individual-level variances for all species ($P < 0.01$), and greater heterogeneity among treatments in replicate-level variances for F ($P < 0.05$) but not for the other two species. Thus, neither models 1 nor 2 adequately characterized differences in variances among treatments.

We derived model 3 (Supplement) to incorporate this heteroscedasticity. Although the parameter estimates from this model were identical to model 2, the precision of their estimators differed, leading to different standard errors, confidence intervals, and P values (Table 1). In contrast to models 1 and 2, model 3 detected statistically significant interspecific competition on F from M (Table 1). Additionally, model 3 gave no statistical indication of interspecific competition on P from M, which was significant in models 1 and 2 (Table 1). Thus, correctly accounting for heteroscedasticity among treatments led to different statistical conclusions.

Non-independence.—All three models detected correlations among measurements from the same basins in those treatments containing heterospecific individuals (interspecific competition treatments). These correlations for (F-M, F-P, M-P) were (0.86, 0.75, 0.98), (0.63, 0.61, 1.00), and (0.91, 0.75, 0.99) for models 1, 2 and 3, respectively. Thus, individuals from both species within the same basin had on average high or low mass. These high correlations, however, had nominal effects on the statistical tests of competition; for model 3, statistically excluding correlations among species from the same replicates did not change the conclusions about the significance of the strengths of competition C (results not presented).

Competition strengths (C) and coefficients (A).—All estimates of A were less than one, implying that interspecific competition was less than intraspecific competition (Table 1). However, the confidence intervals were large and did not exclude 1, which would have indicated estimates of interspecific competition significantly less than intraspecific competition. Also, in comparison to the estimates of the strength of competition C, the estimates and confidence intervals of A differed relatively more between models 1 and 2. This suggests that as estimated parameters involve more-complex combinations of regression coefficients (A vs. C), differences in the estimates and uncertainty of the coefficients are amplified.

Other response variables.—Given heteroscedasticity among treatments for the same species, we used only model 3 to assess competition across the remaining three response variables (Appendix: Table A1). All three species experienced strong intraspecific competition, and the strongest interspecific competition came in the form of reduced survival (logit transformed) of F due to P (Fig. 2). Across all response variables, only two estimates of A were significantly different from 1: reduced stage of M due to F, and reduced survival of F due to P (Table A1).

Discussion

Our goal was to provide a flexible statistical framework to analyze experimental data that contains both complex correlations among treatments and heteroscedasticity. In this approach, all species and all experimental treatments can be included in a single analysis, which facilitates tests of treatment comparisons that are the target of the experiment. This framework should help experimentalists extract the most out of their data.

For our data, model 3 was the best, because it incorporated the heteroscedasticity among treatments that we found. Especially in competition experiments where variation among individuals is expected to differ among treatments, heteroscedasticity should be investigated and, if found, incorporated into the statistical model. This should be part of good statistical practice: performing diagnostics of statistical models to ensure that assumptions are upheld. Presence of heteroscedasticity is generally an indication of biological processes occurring and is not a cause for alarm (Cleasby and Nakagawa 2011). In our case, heteroscedasticity was likely due to competition: the ecological process we were studying. Zuur et al. (2009) illustrate a method for correcting for heteroscedasticity that assumes heteroscedasticity is a function of an explanatory variable. In our case, heteroscedasticity was not a function of any variable and varied among discrete treatments, and therefore this method could not be directly applied. The method we derived for incorporating differences in variances among treatments is not complicated, although it does require some custom computer coding to implement (Supplement).

The technically correct model, model 3, gave different conclusions about competition than the other two models. The three models agreed that all three tadpole species experienced significant intraspecific competition for at least one response variable, and all three species experienced and exerted interspecific competition. However the exact nature of interspecific competition differed between the models, which could lead to different interpretations regarding coexistence. In model 3, P has a strong effect on F, but F has no reciprocal effect on P. This suggests that there is a chance that P could competitively exclude F if the two co-occurred in the same ephemeral pools. In contrast, models 1 and 2 show that F does have an effect on P, which may make coexistence of the two species more likely. There are also discrepancies between the models for interactions between M and F; in model 3 M and F both affect each other, while in models 1 and 2, F only affects M, implying coexistence is less likely. These differences between models illustrate the potential for making incorrect ecological conclusions if the incorrect statistical model is used.

Considering all of our response variables (mass, length, development stage, and survival), there were strong interactions among species according to model 3 (Fig. 2), and these are related to the ecology of the species. F and P have similar trophic ecology and likely compete directly for food. P emerges from eggs larger and earlier than F, which is likely why survival of F was lowered by competition with P while the converse was not true. Asymmetric competition, like that between F and P, is frequently found among anuran species (Morin and Johnson 1988, Parris and Semlitsch 1998, Smith et al. 2004) and broadly across different ecological communities (Lawton and Hassell 1981, Connell 1983, Schoener 1983). This strong survival competition on F from P was also a source of heteroscedasticity in our data. In two blocks of the F + P treatment, we had very low survival of F individuals; however, the survivors (one in one block, five in the other) reached relatively large size. In the other two blocks, survival of F was higher, yet body size variation was also very high. Competition-induced increases in body size variation have been demonstrated in other aquatic systems (Rubenstein 1981, Peacor and Pfister 2006). M experienced strong intraspecific competition and strong interspecific competition from both F and P, measured in terms of mass, length, and development stage. M is a small-bodied filter feeder that has a different feeding niche from the other two species. Nonetheless, it appears to be sensitive to interspecific competition. The large effects of interspecific competition on M did not include reduced survival, however, which suggests plasticity in life-history traits that may allow M to survive in ecologically diverse environments.

Although multilevel models make it possible to perform analyses at the individual level, this is generally not needed for experiments in which individuals are grouped into replicates within a treatment, as in our experiment. We have encountered the recommendation that individual-level models be used to take full advantage of the much larger individual-level sample size, although this sample size is illusory. For experimental data, the true sample size is the number of replicates, and the nearly identical results from our models 1 and 2 illustrate this. However, if our experiment involved repeated measurements on the same individuals over time, such as initial and final masses for each tadpole, then modeling at the individual level might be appropriate. In this case, it would be possible to extract information such as growth rates that would be valuable to compare among individuals within the same treatment.

Finally, we emphasize the value of bootstrapping for hypothesis testing. In addition to giving robust inference for data sets with small sample sizes, bootstrapping is also an effective tool for identifying possible complexities in the data set. For example, we were not convinced of heteroscedasticity until performing the bootstrapping simulations to assess the variance among replicates within treatments (Fig. 1). If a statistical model fits the data, then simulating the model should generate data sets with the same characteristics as the original, and comparing the two gives an easy way to perform diagnostics for complex multilevel models.

Good experiments and associated analyses are needed in ecology. Although we illustrated our framework using a data set from a competition experiment, the framework can be applied to experiments investigating other species interactions, such as predator-prey or mutualisms where interaction coefficients need to be quantified. For example, in an experiment comparing the effectiveness of single and multiple predators on prey suppression (e.g., Straub and Snyder 2006) our framework could be applied to allow the direct comparison between predators, and comparisons of the same predators in the presence/absence of other predators. The statistical plan of setting up treatment contrasts to test complex hypotheses is often useful in experimental analyses, and bootstrapping provides a simple and robust method for statistical inference. More broadly, the issue of heteroscedasticity is likely to be pervasive in ecology, yet we suspect that diagnostics for heteroscedasticity are not routinely performed especially for multilevel models; because multilevel models "build in" the variance structure of the data, researchers may be less likely to make traditional residual plots that would reveal heteroscedasticity. Careful construction of statistical models can not only guard against incorrect statistical conclusions, it can also extract more and more-useful information from data sets.

Acknowledgments

We thank the Ives Lab at UW–Madison and the EEPAI group at XTBG for scientific and logistic support for this work. Brenna Knaebe and Matthew Helmus provided assistance with the experiments. This work was supported by a Fulbright IIE and UW-Madison Zoology Department grant to J. E. Behm and NSF Grant 384 No. DGE-0549369 IGERT: Biodiversity Conservation and Sustainable Development in Southwest China.

Literature Cited

Altwegg, R., and H. U. Reyer. 2003. Patterns of natural selection on size at metamorphosis in water frogs. Evolution 57:872–882.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24:127–135.

Clark, J. S. 2007. Models for ecological data. Princeton University Press, Princeton, New Jersey, USA.

Cleasby, I. R., and S. Nakagawa. 2011. Neglected biological patterns in the residuals: a behavioural ecologist's guide to co-operating with heteroscedasticity. Behavioral Ecology and Sociobiology 65:2361–2372.

Connell, J. H. 1983. On the prevalence and relative importance of interspecific competition—evidence from field experiments. American Naturalist 122:661–696.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York, New York, USA.

Gelman, A., and J. Hill. 2008. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.

Gosner, K. L. 1960. A simplified table for staging anuran embryos and larvae with notes on identification. Herpetologica 16:183–190.

Hayes, A. F., and L. Cai. 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation. Behavioral Research Methods 39:709–722.

Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. Biometrical Journal 50:346–363.

Inouye, B. D. 2001. Response surface experimental designs for investigating interspecific competition. Ecology 82:2696–2706.

Lawton, J. H., and M. P. Hassell. 1981. Asymmetrical competition in insects. Nature 289:793–795.

McCarthy, M. A. 2007. Bayesian methods in ecology. Cambridge University Press, New York, New York, USA.

Morin, P. J., and E. A. Johnson. 1988. Experimental studies of asymmetric competition among anurans. Oikos 53:398–407.

Parris, M. J., and R. D. Semlitsch. 1998. Asymmetric competition in larval amphibian communities: conservation implications for the northern crawfish frog, *Rana areolata circulosa*. Oecologia 116:219–226.

Peacor, S. D., and C. A. Pfister. 2006. Experimental and model analyses of the effects of competition on individual size variation in wood frog (*Rana sylvatica*) tadpoles. Journal of Animal Ecology 75:990–999.

Qian, S. S., and Z. Shen. 2007. Ecological applications of multilevel analysis of variance. Ecology 88:2489–2495.

R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Rubenstein, D. I. 1981. Individual variation and competition in the everglades pygmy sunfish. Journal of Animal Ecology 50: 337–350.

Schoener, T. W. 1983. Field experiments on interspecific competition. American Naturalist 122:240–285.

Smith, G. R., H. A. Dingfelder, and D. A. Vaala. 2004. Asymmetric competition between *Rana clamitans* and *Hyla versicolor* tadpoles. Oikos 105:626–632.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. W. H. Freeman, San Francisco, California, USA.

Straub, C. S., and W. E. Snyder. 2006. Species identity dominates the relationship between predator biodiversity and herbivore suppression. Ecology 87:277–282.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed effects models and extensions in ecology with R. Springer Science + Business Media, New York, New York, USA.

Supplemental Material

Appendix

Strengths of intra- and interspecific competition (C) and competition coefficients (A) for all four response variables (mass, length, developmental stage, and survival) (*Ecological Archives* E094-134-A1).

Supplement

Competition experiment data set and R code for model 1 (individual level), model 2 (basin level), and model 3 (basin level accounting for heteroscedasticity) (*Ecological Archives* E094-134-S1).