Animal Behaviour 79 (2010) 439–450

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

A method for exploring the structure of behavioural syndromes to allow formal comparison within and between data sets

Niels J. Dingemanse [a,*,3], Ned Dochtermann [b,1,3], Jonathan Wright [c,2]

[a] Animal Ecology Group, Centre for Ecological and Evolutionary Studies & Department of Behavioural Biology, Centre for Behaviour and Neurosciences, University of Groningen, The Netherlands
[b] Department of Biology, University of Nevada
[c] Institute of Biology, Norwegian University of Science and Technology

ARTICLE INFO

Article history:
Received 14 April 2009
Initial acceptance 26 June 2009
Final acceptance 20 October 2009
Available online 22 December 2009
MS. number: 09-00260R

Keywords:
behavioural syndrome
common principal components analysis
personality
population differentiation
structural equation modelling
variance–covariance matrix

Research on behavioural syndromes (consistent individual differences in suites of correlated behaviours) requires formal statistical methods to describe and compare syndrome structures. We detail the shortcomings of current methods aimed at describing variation in behavioural syndromes, such as multiple pairwise correlations and principal components analysis (PCA). In their place we propose an alternative statistical framework involving: (1) calculation of trait variance–covariance and correlation matrices within each data set; (2) statistical evaluation of specific hypotheses regarding how behaviours covary within a behavioural syndrome; and (3) statistical comparison of behavioural covariances across data sets using structural equation modelling (SEM). Given their unfamiliarity to most behavioural ecologists, we illustrate these methods using an already published data set for two groups of populations of three-spined stickleback, *Gasterosteus aculeatus*, living in ponds with and without fish predators. Previous analyses suggested a lack of behavioural syndrome structure for stickleback that lived in the absence of fish predators. However, by evaluating a priori hypotheses of how behaviours might covary using SEM, we were able to demonstrate that the two types of populations differed specifically in covariance patterns for aggression, exploration of novel food sources and altered environments, but not for exploration of novel environments and activity. Such detailed inferences cannot readily be made based on conventional statistical approaches alone, and so the methods we outline here should become standard in studies concerning the evolution of behavioural syndromes within and between populations.

The recent discovery that consistent individual differences in suites of correlated behaviours, so-called 'behavioural syndromes' (Clark & Ehlinger 1987; Sih et al. 2004) or 'animal personality' (Gosling 2001; Réale et al. 2007), are common across taxa has led to a sharp increase in the number of studies documenting patterns of behavioural variation between individuals of the same population (Réale et al. 2007; Biro & Stamps 2008; Sih & Bell 2008; Dingemanse et al., in press). For example, individuals that are relatively aggressive are often also relatively bold compared to less aggressive individuals in their population (Réale et al. 2007).

Evolutionary theory predicts that trait correlations, like the ones that define behavioural syndromes (Bell 2007), can evolve in response to selection (Arnold 1992), implying that syndrome structure could differ between selective environments or populations depending on their evolutionary history of selection. For example, studies on three-spined stickleback, *Gasterosteus aculeatus*, have revealed that correlations between aggressiveness and boldness are stronger in populations where fish predators are present (versus absent); fish predators potentially induce selection which favours correlations between behavioural traits (Bell 2005; Bell & Sih 2007; Dingemanse et al. 2007). In other study systems it might well be that behavioural syndromes are ubiquitous, either because natural selection induces correlational selection in any type of population (Dingemanse & Réale 2005) or because strong pleiotropic effects of hormones or genes have resulted in syndromes that cannot easily be modified in the course of evolution ('constraint hypothesis'; Bell 2005). Therefore, at some point all

* Correspondence and present address: N. J. Dingemanse, Department of Behavioural Ecology & Evolutionary Genetics, Max Planck Institute for Ornithology, Seewiesen, Germany.
E-mail address: ndingemanse@orn.mpg.de (N.J. Dingemanse).
[1] N. Dochtermann is at the Department of Biology, MS314, University of Nevada, Reno, NV 89557, U.S.A.
[2] J. Wright is at the Institute of Biology, Norwegian University of Science and Technology, 7491 Trondheim, Norway.
[3] These authors contributed in equal part to this work.

0003-3472/$38.00
doi:10.1016/j.anbehav.2009.11.024

researchers studying behavioural syndromes will have to decide how to evaluate variation in syndrome structure and assess the explanatory power of different hypothesized syndrome structures (Dochtermann & Jenkins 2007).

In this paper, we discuss the methods that have been applied to describe both behavioural syndromes and variation in syndrome structure. We outline advantages and shortcomings of each method and we propose an alternative statistical framework for the analysis of behavioural syndrome variation. We illustrate the proposed framework by reanalysing published data describing different syndromes from populations of three-spined stickleback. Our previous analyses of these data showed that behavioural syndromes were present in six populations living sympatrically with predators, compared with another six populations living without predators where syndrome structure appeared to be totally absent (Dingemanse et al. 2007). We describe here the novel insights regarding similarities and differences in syndrome structure across our stickleback populations based on the application of the proposed statistical framework, thereby illustrating the added value of the advocated approach.

METHODS FOR DESCRIBING SYNDROME STRUCTURE

Before outlining a new methodology for studying behavioural syndrome structure, we should emphasize that correlations between two behavioural traits can exist both between and within individuals. The former represents a behavioural syndrome, whereas the latter type represents the presence of 'correlational plasticity' within individuals (i.e. within-individual changes in one behaviour from one observation to the next correlate with changes in another behaviour Dingemanse et al., in press). The interindividual correlation that defines a behavioural syndrome can readily be documented by subjecting the same set of individuals repeatedly to two (or more) independent behavioural assays. For instance, to study the presence of an aggressiveness – boldness syndrome, one could design separate assays for measuring baseline activity, aggression, and boldness (Huntingford 1976). Researchers should then summarize the variation in all the behavioural elements measured within each of the assays into a composite (assay-specific) 'score' in such a way that the phenotypic variance in each composite score best reflects the interindividual variance in the behaviour of interest (for guidelines see Dingemanse et al., in press).

The majority of published studies estimate the sign, strength and significance of the correlation between each of the (two or more) assayed behaviours or composite scores. For example, Bell (2005) analysed behavioural syndromes in two stickleback populations by calculating the correlation between each of three behaviours (activity, aggression and boldness). Results are then interpreted in the context of alternative hypotheses for syndrome structure (for empirical examples see studies reviewed in Réale et al. 2007; Sih & Bell 2008). Alternative hypotheses range from the possibility that behaviours vary independently of one another to the possibility that all behaviours covary similarly in all situations (see Introduction).

The advantage of this approach is the simple nature of the statistical methodology: when two behaviours are significantly correlated, we can conclude that there is statistical evidence for the presence of a syndrome. However, there are a number of disadvantages inherent in this approach. First, the pairwise correlations between all behaviours that one can calculate for a given set of individuals are, by definition, not independent (Dietz 1983). Thus, conclusions based on separate analyses of each pairwise correlation cannot be drawn with much confidence. This lack of confidence stems from an increased probability of type I statistical errors caused by multiple testing (Quinn & Keoug 2002). Second, the statistical power of each pairwise correlation might be low if type I error rates for multiple testing are controlled for; implying an increased probability of type II statistical errors in cases where sample sizes are low or behavioural syndromes are present but weak (Dochtermann & Jenkins 2007). Third, bivariate correlations might not properly represent the true covariance between two traits, because they ignore combined effects that can be more properly characterized by the use of partial correlation coefficients. Fourth, this type of methodology does not provide objective information on how well the overall data set fits any specific hypothesis concerning syndrome structure, implying that one cannot easily formulate objective criteria by which successive hypotheses are accepted or rejected.

Other methods that have been used to document behavioural syndromes are factor analysis and principal components analysis (Mather & Anderson 1993; Gosling & John 1999; Gosling 2001). Such analyses are carried out using all the assay-specific behavioural measures or composite scores together to work out which ones are correlated (i.e. a syndrome) and whether all these correlations can effectively be captured in a single standardized principal component (PC) or multiple orthogonal PCs, which effectively describe the behavioural syndrome. Although apparently more complex than performing simple correlations, these data reduction methods avoid the issues of type I and type II errors, because no hypothesis is tested at all. Unfortunately, this type of methodology therefore cannot provide objective information on how well the overall data set fits specific hypotheses for presumed syndrome structure.

An alternative method for syndrome structure analysis has recently been proposed by Dochtermann & Jenkins (2007). Using confirmatory factor analysis implemented as structural equation modelling (SEM), these authors calculated how well alternative hypotheses for syndrome structure fit all the available behavioural data. The main advantage of this methodology is the ability to compare statistically the fit of alternative hypotheses of syndrome structure (Dochtermann & Jenkins 2007). The use of SEM also allows the strength and direction of relationships between traits to be evaluated while controlling for covariance between other traits. This statistical control can reveal patterns of phenotypic trait covariance not apparent when just bivariate correlations are used, as shown in our worked example below.

METHODS FOR COMPARING SYNDROME STRUCTURE ACROSS DATA SETS

To understand the evolution and potentially adaptive nature of behavioural syndromes, we need to be able to compare syndrome structure between data sets (e.g. experimental treatments), populations and potentially between species formally. To date, relatively few studies have documented syndrome structure in more than one data set (Cade & Cade 1992; Riechert & Hedrick 1993; Bell & Stamps 2004; Bell 2005; Bell & Sih 2007; Dingemanse et al. 2007, 2009; Moretz et al. 2007; Brydges et al. 2008; Sinn et al. 2008), and even fewer have statistically determined whether syndrome structures are equivalent between data sets.

To compare syndromes across data sets, researchers can estimate the sign, strength and significance of the correlations between each of the assayed behaviours, and then test whether each focal correlation differs significantly between data sets (Bell 2005; Bell & Sih 2007; Dingemanse et al. 2007, 2009). These methods have the same advantages and disadvantages when applied across data sets as when implemented for a single data set as described above. Because of these limitations, multivariate approaches are a preferable alternative for comparing syndrome structures across data sets. Multivariate statistical techniques to

evaluate differences in trait associations across populations are frequently used by evolutionary biologists (Arnold & Phillips 1999; Steppan et al. 2002) to test whether the variance (usually additive genetic variance) and correlations (usually additive genetic correlations) differ between populations or treatments. However, these approaches have rarely been used in the context of behavioural syndrome research (but see Dingemanse et al. 2009). This general absence of multivariate approaches is perhaps due to the relatively recent emergence of behavioural syndrome research.

Multivariate methods summarize the variances and covariances for all behaviours or composite scores that may contribute to the syndrome in variance–covariance matrices. Variation in syndrome structure can then be studied by statistically comparing variance–covariance matrices across data sets. One common method for comparing matrices is common principal component analyses (CPC; Phillips & Arnold 1999). Using CPC, covariance matrices can be shown to be: (1) identical (implying that all (co)variances are the same); (2) proportional (implying that data sets differ in the amount of variation in individual behaviours, but not in the covariation between behaviours); or (3) nonproportional (at least one covariation between behaviours differs between data sets, i.e. differences in syndrome structure). Unfortunately, the CPC-method only provides a rough insight into any (dis)similarity between matrices and its use has been criticized because biological interpretation of matrix dissimilarity is not straightforward (Blows 2007; Brodie & McGlothlin 2007). Other methods for the comparison of variance–covariance matrices include Mantel tests (Lofsvold 1986; Kohn & Atchley 1988), modified matrix element comparisons (Roff et al. 1999) and maximum-likelihood comparisons (Shaw 1991). However each of these approaches has a variety of limitations (Shaw 1991; Phillips & Arnold 1999; Roff 2002).

THE PROPOSED FRAMEWORK

Given the pros and cons involved with each of the methods mentioned above, we propose the following framework for the analyses of syndrome structure within data sets, and for the comparison of syndrome similarity between data sets: first, we recommend the calculation of variances, covariances and correlations for each data set, such that (dis)similarities in each of these parameters can be qualitatively compared. Second, we recommend comparison of a priori formulated alternative models (hypotheses) for syndrome structure for each data set separately and the statistical evaluation of relative fit of each model (Dochtermann & Jenkins 2007). Finally, we recommend statistical comparison of structural models across data sets for detailed insight of the quantitative (dis)similarity of behavioural variance–covariance matrices. This comparison can be conducted using 'multigroup analysis' where covariance patterns (e.g. the direction of correlations) are compared between groups. The major advantage of this three-part approach is that it allows the specific testing of hypothesized patterns of behavioural covariance. The quantitative comparison between groups also creates the possibility that evolutionary hypotheses regarding behavioural syndromes can be evaluated directly. Neither of these advantages is conferred by currently used approaches and neither the combination of these three steps nor multigroup analyses in particular has been conducted within the context of behavioural syndrome research.

A WORKED EXAMPLE

The Data

We demonstrate the utility of the proposed framework by reanalysing data previously reported by Dingemanse et al. (2007) from 12 populations of three-spined stickleback, with seven individuals per population (84 individuals in total). Six of the 12 populations were from larger lakes where the stickleback occurred sympatrically with fish predators (PRED populations). The other six populations were from smaller ponds that did not contain fish predators (NAÏVE populations). Five behaviours were assayed for each individual within the categories of aggression (towards conspecifics), general activity (in a familiar environment) and exploration–avoidance (of novel foods, novel environments and altered environments). For more details see Dingemanse et al. (2007). We controlled for environmental effects that could potentially bias the data (e.g. time of day, season) by estimating best linear unbiased predictors (BLUPs) for behaviours of each individual as recommended by Martin & Réale (2008). BLUPs can be considered as analogous to residuals from an analysis of variance or regression and in this case they represent individual behavioural responses. Based on these data, Dingemanse et al. (2007) demonstrated behavioural syndrome structure in PRED but not NAÏVE populations.

Statistical Analyses

The proposed analyses could not be conducted with a sample size of seven individuals per population. We therefore pooled the data into two larger data sets, one for each type of population (PRED and NAÏVE), after *z*-transforming behavioural responses for each population separately. This removed population differences in mean and variance within each type of population while maintaining patterns of covariance. We then applied the three-part statistical analysis of syndrome structure:

(1) Descriptive statistics. For each type of population we calculated behavioural variances and covariances along with their associated Pearson correlations for all pairs of behaviours. Variances and covariances and the significance of correlations were calculated using SPSS version 14.0 (SPSS Inc., Chicago, IL, U.S.A.) and population-specific values of trait variances are available in Dingemanse et al. (2007).

(2) Structural equation modelling within data sets. We formulated five a priori hypotheses relating to syndrome structure as specified below. For each type of population we constructed a set of factor models implemented as structural equation models (one for each hypothesis, using AMOS 7.0, SPSS Inc.) and statistically compared the different models using the Akaike's information criterion (AIC). AIC values were calculated from model discrepancies ($\hat{C}$) estimated by maximum likelihood using bootstrapping (with 1000 bootstraps). AIC values balance the fit of a model to the data while penalizing for complexity and thus rewarding parsimony with lower values indicating greater support for a particular statistical model and its corresponding hypothesis (Akaike 1973; Burnham & Anderson 2002). AIC values were evaluated based on AIC differences relative to the model with the lowest AIC (ΔAIC). ΔAIC values greater than two suggest decreased support for a particular model of syndrome structure relative to the model with the lowest AIC value. ΔAIC values greater than 2 do not imply 'no support' or the falsification of a competing hypothesis, but simply indicate less support (Burnham & Anderson 2002).

Five a priori hypotheses of syndrome structure were considered based on the behavioural syndrome literature (models 1–5, Fig. 1). Model 1 represented the absence of covariance, with behavioural responses varying independently of one another (Coleman & Wilson 1998). Models 2 and 4 represented domain-general models of syndrome structure (Sih et al. 2004) with exploratory behaviour, activity and aggression linked via some underlying factor. Models 2 and 4 differ depending on whether exploration of food is considered a contextually different behavioural response compared to

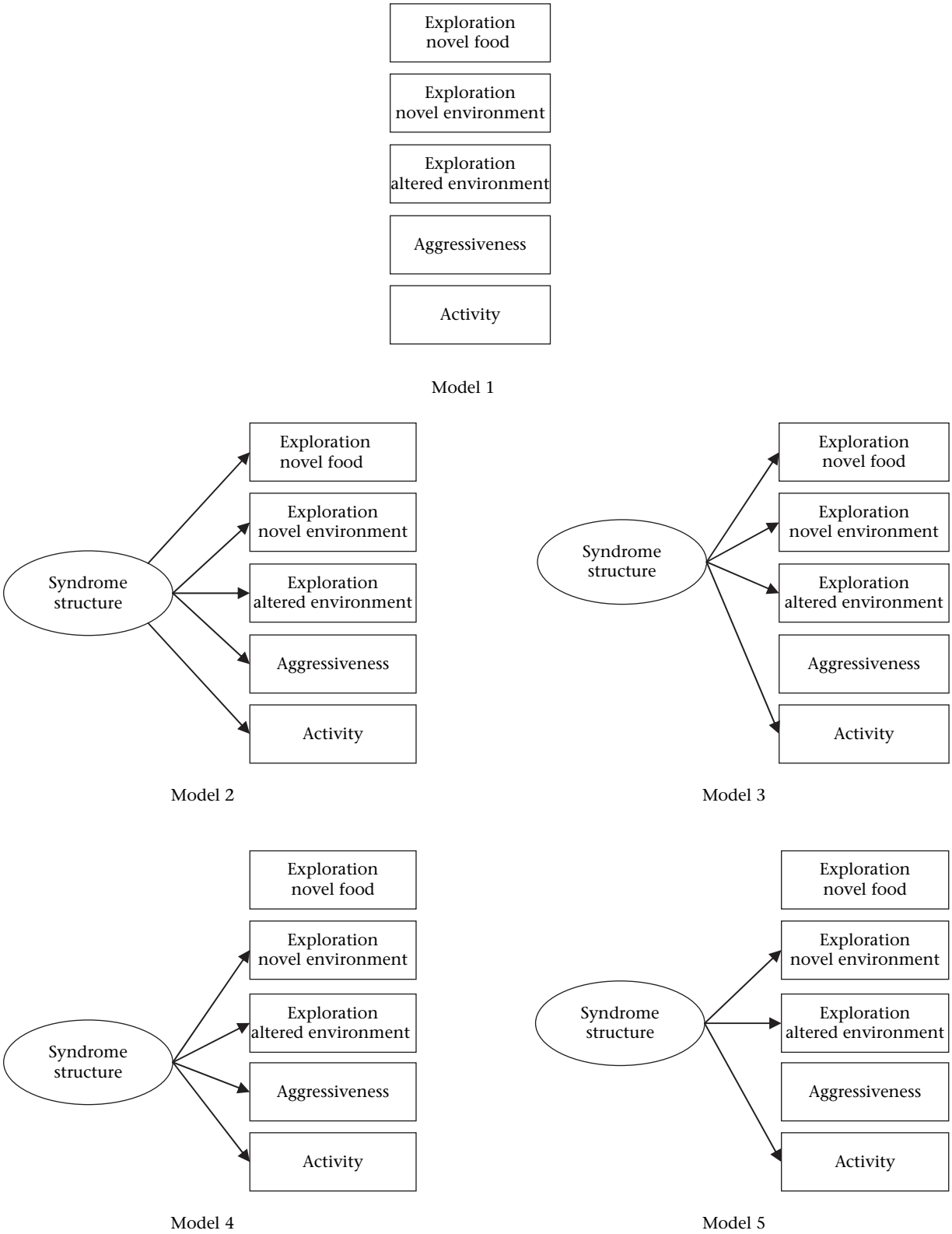

Figure 1. Five models of syndrome structure developed based on a priori hypotheses of behavioural syndrome structure. Model 1 represents behavioural independence, which may be evolutionarily advantageous (Coleman & Wilson 1998). Models 2 and 4 are domain-general models of syndrome structure while models 3 and 5 are more restricted domain-general models with only exploration and activity covarying (Sih et al. 2004). The measured behaviours are represented in rectangular boxes. Underlying causal connections (latent variables) resulting in syndrome structure are represented in ovals (Grace 2006).

other exploratory activities (Coleman & Wilson 1998; Mettke-Hofmann et al. 2002, 2005; Mettke-Hofmann 2007). Models 3 and 5 represented more restricted domain-general models of syndrome structure (Sih et al. 2004) with exploration and activity linked but with aggression varying independently. Models 3 and 5 differed depending on the contextual basis of food exploration (Coleman & Wilson 1998; Mettke-Hofmann et al. 2002, 2005; Mettke-Hofmann 2007). In models 2–5 (Fig. 1) 'syndrome structure' is statistically modelled as what is known as a 'latent variable' (Grace 2006).

Because only a priori hypotheses are considered in this analysis it is possible that no hypothesis explains a biologically substantive amount of variation (Dochtermann & Jenkins, in press). To assess this possibility, the proportion of the variation (R^2) in each behaviour explained by the syndrome structure itself can be calculated (Tabachnick & Fidell 2001). D_x, an analogue for the R^2 value, for the entire set of behaviours can also be calculated. D_x represents the proportion of variation in the behavioural variance–covariance matrix that is explained by a particular model of syndrome structure, relative to no syndrome structure (Stamps et al. 2005). R^2 values are readily calculated by most SEM statistical packages for each behaviour. The overall D_x is calculated as:

$$D_x = 1 - \frac{\hat{C}_x}{\hat{C}_{\text{null}}}$$

where $\hat{C}_{\text{null}}$ is the estimated discrepancy for a model of no behavioural syndrome (i.e. model 1, Fig. 1) and $\hat{C}_x$ is the estimated discrepancy for an alternative model (in our case models 2–5, Fig. 1). D_x can be interpreted in the same manner as an R^2 value. While overall R^2 values can be calculated for confirmatory factor analyses (e.g. Tabachnick & Fidell 2001), D_x is preferable here because of how null expectations are estimated and because all the models are evaluated against the model of no syndrome structure.

(3) Comparison of SEM models across data sets. We statistically compared SEM models between the two groups for detailed insight concerning quantitative (dis)similarity of behavioural matrix elements using a two-part multigroup analysis (Grace 2006). For the first portion of this multigroup analysis, statistical comparisons were conducted between the PRED and NAÏVE groups to evaluate whether phenotypic traits exhibited similar patterns of covariance. This multigroup analysis is qualitatively similar to the CPC analyses mentioned earlier.

In a multigroup analysis, the fit of two or more data sets to a particular model are compared in a hierarchical manner. First, one compares whether or not there is the same syndrome structure between the groups; that is, are the same behaviours connected within a syndrome structure between the data sets? For example, do behaviours for sticklebacks in our NAÏVE group conform to a behavioural syndrome as hypothesized in model 3 (Fig. 1) while sticklebacks in the PRED group instead have a behavioural syndrome like model 4 (Fig. 1)? Second, if the structures are the same, one compares the strength and direction of the relationships between 'syndrome structure' and specific behaviours and whether this differs between groups (often termed H_λ). Third, one compares whether the variance in individual behaviours differs between groups. These three comparisons are evaluated between groups based on goodness-of-fit tests and the degrees of freedom determined by how many variables would be the same between groups. If the first comparison, the test of similar syndrome structure, is statistically supported then behavioural syndromes connect the same behaviours in the different groups. If the second comparison, the test of the strength and direction of relationships, is supported then behavioural syndrome structures connect behaviours in the same way between the different groups. If the third comparison is statistically supported then behaviours vary equally between groups. Multigroup analyses testing these hypotheses can be conducted using most SEM software packages.

Based on stickleback syndrome data from a set of independent studies (Bell 2005; Bell & Sih 2007), we predicted that NAÏVE and PRED groups would have different behavioural syndrome structures. We expected that the NAÏVE grouping of populations would have a syndrome structure consistent with either structural equation model 3 or model 5, while the PRED grouping of populations would have a syndrome structure consistent with either model 2 or 4 (see Fig. 1). Thus, we predicted that the multigroup analysis would reveal that the groups would not have syndrome structures of the same form (step 1 of the hierarchy described above).

For the second portion of the multigroup analysis, differences between models in how behaviours are causally related to underlying factors can be determined. This portion of the analysis should only be conducted if the second step of hierarchical comparisons (H_λ) is not statistically supported but the first is. The comparison of relationships (i.e. a comparison of factor loadings) can be conducted using a multivariate Lagrange multiplier-based method with specialized software (Grace & Jutila 1999). Alternatively, loadings can be compared using standard approaches for comparing regression coefficients (Zar 1999). Using this second approach, the difference between loadings is determined along with the pooled standard error for each path (determined by analysing all the data together). The ratio of the difference to the standard error is then evaluated as a t value to determine significance (Zar 1999). Degrees of freedom are based on the sum of the degrees of freedom for each group (Zar 1999).

As an alternative to a formal multigroup analysis, researchers can also include experimental groups or populations as a factor in the SEM model. For example, 'presence of predators' could be included as an observed variable (recorded as no predators, '0', or with predators, '1') in our models with direct effects on 'syndrome structure' (Fig. 1, models 2–5). This approach could be considered similar to an analysis examining interaction effects and the significance of the loading between the grouping variable and 'syndrome structure' could then determine whether there are significant group differences. Different models incorporating this grouping variable could also be compared. However, this approach would not identify what the specific differences between groups would be. It would be akin to having a significant 'treatment' effect in an analysis of variance but not conducting post hoc contrasts. In contrast, the comparison of loadings as regression coefficients can be considered a post hoc assessment of what specific behaviours differ between groups.

RESULTS AND DISCUSSION

Descriptive Statistics

Calculation of pairwise phenotypic correlations showed very little evidence for syndrome structure in the NAÏVE group: three behaviours (activity, exploration of novel environments and of altered environments) were significantly correlated, but only one of those correlations (activity versus exploration of altered environment) remained significant when we used a sequential Bonferroni adjustment to control for familywise error rates (Table 1). In contrast, four behaviours (activity, aggression, exploration of novel and of altered environments) were significantly correlated in the PRED group, and all of those, except the correlation between aggression and exploration of novel environments, remained so after sequential Bonferroni correction (Table 1). While Bonferroni corrections are often considered overly conservative (Garcia 2004; Garamszegi 2006), correction is needed unless the results are to be considered purely exploratory.

Table 1
Phenotypic correlations between five behavioural traits for three-spined stickleback

Behaviour†	Activity	Aggression	Exploration of		
			Novel environment	Novel food	Altered environment
Activity	—	−0.01	0.43**	0.09	**0.49****
Aggression	0.24	—	−0.27	0.11	−0.13
Exploration of novel environment	**0.61*****	0.33*	—	−0.06	0.33*
Exploration of novel food	0.03	0.21	0.04	—	0.10
Exploration of altered environment	**0.71*****	**0.48*****	**0.75*****	0.06	—

Data above the diagonal are for the predator-naïve group, below the diagonal for the predator-sympatric group.
† Values of P do not control for multiple testing of the same data (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). Only values printed in bold are significant after sequential Bonferonni adjustment of experimental error rates (Zar 1999).

Structural Equation Modelling within Data Sets

Contrary to our predictions, and the descriptive statistics reported above, both NAÏVE and PRED groups had a syndrome structure in which aggression covaried with exploratory behaviours and activity (Table 2). For the NAÏVE group, the model that best explained the data was one in which exploration–avoidance of a novel food covaried with aggression, activity and other exploration–avoidance behaviours (model 2; Table 2). For the PRED group there was equal support for models in which exploration–avoidance of a novel food covaried with (model 2, $\Delta AIC = 0.5$; Table 2) or varied independently of other syndrome components (model 4, $\Delta AIC = 0$; Table 2).

The best models explained a major portion of the variation present in behaviour variance–covariance matrices for both groups. For the NAÏVE group, a 'model 2' hypothesis of trait covariance explained approximately 77% of the variance–covariance matrix variation in behaviour (Table 2). For the PRED group we could not statistically distinguish between the 'model 2' and 'model 4' hypotheses of trait covariance, because both explained approximately 86% of the variance–covariance matrix variation in behaviour (Table 2).

Comparison of SEM Models Across Data Sets

SEM analyses supported domain-generality in syndrome structure in both types of population (i.e. support for model 2 in the *NAÏVE* group, or model 2 and 4 in the PRED group, Table 2).

Table 2
Model comparison results for predator-naïve and predator-sympatric groups

Model (x)	$\hat{C}$ (discrepancy)	k	AIC	ΔAIC	Model weight	D_x
Predator-naïve						
2	8.05	10.00	28.05	0.00	0.61	0.77
4	12.44	9.00	30.44	2.39	0.19	0.64
3	12.72	9.00	30.72	2.67	0.16	0.63
5	17.45	8.00	33.45	5.41	0.04	0.50
1	34.79	5.00	44.79	16.74	0.00	0.00
Predator-sympatric						
4	12.03	9.00	30.03	0.00	0.56	0.86
2	10.53	10.00	30.53	0.50	0.44	0.87
5	28.72	8.00	44.72	14.69	0.00	0.66
3	27.37	9.00	45.37	15.34	0.00	0.67
1	83.86	5.00	93.86	63.83	0.00	0.00

Structural equation models (SEMs) were evaluated based on difference in Akaike's information criterion (AIC) values, with small values indicating a better parsimony-informed fit to the data. AIC values were calculated based on the discrepancy between the statistical model for a hypothesis ($\hat{C}$) and the number of parameters (k). D_x values represent the proportion of the variance explained by the focal model relative to null expectations of no syndrome structure. D_x can be considered analogous to R^2.

However, such evidence does not provide information regarding the statistical (quantitative) similarity of syndrome structure for the two groups. Therefore, using 'model 2' for both NAÏVE and PRED groups (Fig. 2), we conducted a multigroup analysis (Grace 2006) to determine whether the same structural equation model provided a fit for both *NAÏVE* and PRED groups. The multigroup analysis demonstrated that NAÏVE and PRED groups did not share similar path coefficients ($\chi^2_{14} = 26.6$, $P = 0.021$; H_λ not supported).

We next compared the strength of the connection between the underlying syndrome structure and individual behaviours to clarify further the differences between NAÏVE and PRED groups. These strengths, or loadings, were compared between the groups using standard approaches for the comparison of regression coefficients (Zar 1999). This analysis showed that loadings differed significantly between NAÏVE and PRED groups for aggression, exploration of novel food and exploration of altered environments (Table 3).

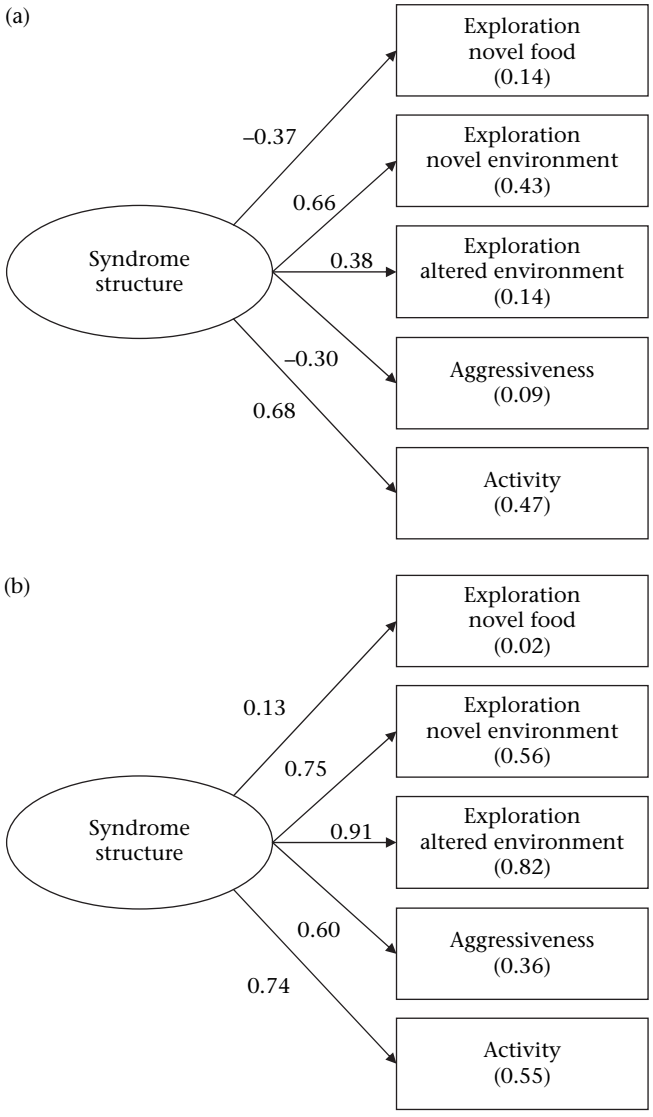

Figure 2. Structural equation models (SEMs) showing how behaviours were related within (a) NAÏVE and (b) PRED groups of stickleback. Numbers in parentheses are the variances of the different behaviours explained by the SEM structure (R^2) for 'model 2' (see Fig. 1). Numbers associated with arrows are standardized factor loadings for the effects of the underlying syndrome structure on a particular behaviour. These represent how behavioural responses are predicted to change based on changes to the underlying syndrome structure (e.g. genetic or hormonal changes). For example in the NAÏVE group, a shift of 1 SD along the distribution of syndrome structures for the population would result in a 0.68 SD increase in activity.

Table 3
Comparison of factor loadings between predator-naïve and predator-sympatric groups based on the 'model 2' hypothesis (domain-generality) of syndrome structure

Path to	Difference	Pooled SE	*t*	*df*	*P*
Exploration of novel food	−0.459	0.116	−3.957	10	**0.003**
Activity	−0.052	0.113	−0.460	10	0.328
Exploration of altered environment	−0.487	0.109	−4.468	10	**0.001**
Exploration of novel environment	−0.085	0.110	−0.773	10	0.229
Aggressiveness	−0.829	0.116	−7.147	10	**<0.001**

The paths are shown to each of the behaviours: exploration of a novel food; activity; exploration of an altered environment; exploration of novel environment; and aggressiveness. We followed Zar (1999) for comparing regression coefficients. Values of *P* are for a two-tailed *t* test. All significant values of *P*, printed in bold, remained significant after sequential Bonferonni adjustment of experimental error rates (Zar 1999).

Loadings of assayed behaviours all had the same sign for the PRED group (Fig. 2b), implying that an as yet unknown proximate factor or factors affected the expression of all assayed behaviours in the same direction (e.g. a gene with pleiotropic effects, or selection-induced linkage disequilibrium). Relatively aggressive individuals were therefore also relatively active and explorative (in all contexts) compared to relatively nonaggressive individuals in this type of population. In contrast, path coefficients of the assayed behaviours did not have the same sign for the NAÏVE group (Fig. 2a). Here, the factor causing the syndrome affected all assayed behaviours, although not all in the same direction: relatively active individuals were also relatively explorative in novel and altered environments (as in the PRED group), but this type of individual was relatively less aggressive and less explorative when confronted with a novel food, which was unexpected.

GENERAL DISCUSSION

Our findings demonstrate the increases in information regarding differences in behavioural syndrome structure that can be obtained with SEM-based confirmatory factor analysis. By evaluating a priori hypotheses as to how behaviours may covary using SEM, and by using multigroup SEM analysis, we demonstrated that PRED and NAÏVE stickleback populations differed in covariance patterns for aggression, exploration of novel food sources and altered environments, but not for exploration of novel environments and activity.

The demonstration of behavioural trait covariance in both NAÏVE and PRED groups is also interesting in its own right, especially given that both the descriptive statistics (based on calculation of bivariate correlations) and the previous statistical analyses of these data (Dingemanse et al. 2007) failed to show strong behavioural covariance in NAÏVE populations. Such discrepancies between univariate approaches (pairwise correlations) and multivariate approaches (SEM) were to be expected, because the former method suffers from increased probability of committing type II statistical errors, underlining the added value of SEM in behavioural syndrome research (Dochtermann & Jenkins 2007, in press).

SEM also demonstrated that the proximate factor(s) underlying the behavioural syndrome in the NAÏVE group had dissimilar effects on behaviours to the PRED group, leading us to conclude that the behaviours aggressiveness and exploration of novel foods were positively associated with one another, but negatively associated with the other behaviours. Behavioural syndrome structure also explained very different amounts of behavioural variation between the two groups. In contrast, the initial bivariate correlations did not reveal this degree of detail with regard to group differences. Again, such discrepancies between methods are to be expected because pairwise correlations do not control for covariance with other traits (Grace 2006), which illustrates that patterns based on SEM can indeed differ in both magnitude and direction from bivariate correlations.

Data Requirements for SEM

Researchers interested in applying the approach advocated in this paper should be aware of certain data requirements for SEM. In short, as with other multivariate approaches, SEM assumes multivariate normality and can require large sample sizes.

Prior to conducting an SEM analysis of behavioural syndrome structure, researchers should assess the assumption of multivariate normality. Determining fit and thus the ability to compare different models of syndrome structure is contingent on multivariate normality (Kline 2005; Grace 2006). While multivariate normality can be difficult to assess, researchers should at least determine whether or not variables are individually normally distributed. If not, variables should be transformed and/or bootstrapping used (Grace 2006). We conducted bootstrapping for the stickleback analyses.

Several diagnostic tests can also be used prior to conducting an SEM-based analysis of behavioural syndrome structure. First, whether or not the behavioural variance–covariance matrix differs from random can be determined using a Bartlett's test of sphericity. If a matrix does not differ from a random expectation then the assertion of syndrome structure may be spurious (S. V. Budaev, unpublished data). In our example, the matrices for both the PRED and NAÏVE groups differed from random ($\chi^2_{10} = 69.2$, $P < 0.01$ and $\chi^2_{10} = 21.3$, $P = 0.019$, respectively). However, it should be noted that Bartlett's test of sphericity has limited power when correlations average below 0.2 (Reddon & Jackson 1984).

As a second diagnostic, since the SEM approach proposed here is an implementation of confirmatory factor analysis, there is another index, the Kaiser–Meyer–Olkin (KMO) index, that can be calculated as an assessment of whether an analysis should proceed. KMO values should be greater than 0.5 if factor analysis is to be conducted (Rencher 2002). In our example, KMO values were greater than 0.5 for both the PRED and NAÏVE groups (KMO = 0.73 and 0.62, respectively). However, because KMO is intended for exploratory factor analyses, its utility for a confirmatory factor analysis with only a single factor, like the one implemented in our worked example, is less clear. Bartlett's test can be conducted in a variety of statistical packages, as can the calculation of KMO values.

The proper implementation of SEM may also be highly dependent on sample size. Several authors have suggested that SEM and confirmatory factor analyses require large sample sizes (e.g. >100). For example, Barrett (2007) asserted that SEM evaluation based on model discrepancy and the rejection of null hypotheses cannot be conducted reliably with sample sizes less than 200. This recommendation was based largely on concerns about low statistical power and the requirement of sample sizes greater than 200 is supported by modelling studies demonstrating that some measures of statistical fit failed to reject models known to be incorrect except at very large sample sizes (Bentler & Yuan 1999). Similarly, Kline (2005) recommended that complex causal models should not be evaluated without sample sizes of at least 100, while Grace (2006) recommended a sample size of 50 as a general minimum.

However, when considering sample size it is important to note that much of the examination of SEM methods has been conducted in the social sciences where very complex models are often considered. Barrett's (2007) recommendation (of not conducting SEM evaluation with sample sizes of less than 200) was in the

context of a model containing dozens of measures. Similarly, Bentler & Yuan's (1999) assessment of different fit statistics was based on a model containing over 30 estimated variables. In contrast, the models we have evaluated here required the estimation of less than one-third of the number of variables. Researchers implementing the framework we have described here are likely to evaluate models of similar complexity.

For the relatively simple models expected in behavioural syndrome research, SEM can be conducted with much smaller sample sizes than those required for the types of complex models used in the social sciences (Kline 2005). Moreover, because researchers investigating behavioural syndrome structure are likely to be dealing with relatively simple models, the ratio of the sample size to the number of variables may be more important than specific recommendations of sample sizes (S. V. Budaev, unpublished data). Ratios of 10 samples per estimated variable are often recommended (Kline 2005). However, Herzog & Boomsma (2009) suggested that estimates of model discrepancy/fit can be adjusted and still be valid even at substantially smaller ratios (e.g. 2:1). Bentler & Yuan (1999) provided a description of the fit statistics that should be used for different sample size:parameter ratios, including guidelines for when sample sizes are less than the available degrees of freedom.

When considering multiple SEMs in competition, the sample size issues discussed above become more complex but less stringent. When using conventional approaches of attempting to reject null hypotheses, if sufficient power is available to reject one model from consideration statistically (based on discrepancy or goodness-of-fit tests) then it may be reasonable to assume that sufficient power exists to evaluate all models under consideration statistically (Violato & Hecker 2007). Consistent with this assertion, the model of behavioural independence (model 1, Fig. 1) in our stickleback example significantly departed from the data for both the PRED ($\chi^2_{10} = 73.66$, $P < 0.01$) and NAÏVE groups ($\chi^2_{10} = 22.65$, $P = 0.01$). However, in the methodology we have described, hypotheses are being evaluated not in competition with null hypotheses but instead in direct competition with one another based on AIC values (Burnham & Anderson 2002).

Ranking models based on AIC values allows a more flexible use of SEM and confirmatory factor analysis. Because models of no relationship between variables can be considered in the analysis (e.g. model 1), concerns about whether the covariance matrix differs from random (e.g. the hypothesis tested by Bartlett's test of sphericity) are evaluated during the ranking procedure. If a model where behaviours vary independently is supported as well as models of syndrome structure, then syndrome structure cannot be inferred.

The sample size requirements discussed earlier were made in the context of conventional approaches for evaluating SEM fit and largely reflect concerns about the power to reject hypotheses of model fit. These issues are far less well defined for a model ranking approach. Haughton et al. (1997) examined how well several information criteria, including AIC as used with the three-spined stickleback data, performed in distinguishing between models. However, this simulation study was conducted based on sample sizes and models inappropriate to our discussion here. Based on the Haughton et al. (1997) framework, we conducted a simulation analysis to examine the performance of AIC values to distinguish between models of syndrome structure at applicable sample sizes (see Appendix). AIC values showed a high power to distinguish properly between models based on these simulations, suggesting that they are particularly appropriate for addressing questions of behavioural syndrome structure.

Nevertheless, because estimates for path coefficients, factor loadings and other SEM parameters are sensitive to variation in sample size, researchers should always maximize the number of observations wherever logistically possible.

Conclusions

The multivariate statistical methodology outlined in this paper has substantial potential when applied to problems in behavioural ecology and specifically to the study of behavioural syndromes. It represents a considerable improvement on current methods and provides greater sensitivity to the exploration and comparison of syndrome structures. In the case of the stickleback data from Dingemanse et al. (2007), it not only formally confirms the presumed differences in syndromes between PRED and NAÏVE populations, but more precisely identifies the exact elements within the syndromes that differ between these two selective regimes. An important reason for applying SEM is the ability to uncover trait associations that would normally remain obscured: The hidden negative covariance between certain behaviours detected in the NAÏVE population might, for instance, suggest that population-specific evolutionary trade-offs between those traits exist in natural populations (Sgro & Hoffmann 2004; Sih et al. 2004), a notion that would not have been prompted had only conventional techniques been used. We would therefore argue that the statistical framework outlined here represents the future for studies of the evolution of behavioural syndromes.

Acknowledgments

We thank Sergey Budaev and an anonymous referee for very insightful and helpful comments on the manuscript. These comments greatly improved the clarity of this paper. We also thank Dawn Thomas, Nick Dawnay, Rachael Hickling and Ashely Tweedale for assistance during fieldwork, and the Countryside Council for Wales, Welsh Water, the Royal Society for the Protection of Birds, and private landowners of the 12 stickleback ponds for permission to capture stickleback on their properties. N.J.D. was supported by the Netherlands Organisation for Scientific Research (grants S 84–566 and 863·05·002).

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: *International Symposium on Information Theory* (Ed. by B. N. Petran & F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadi.

Angilletta, M. J., Oufiero, C. E. & Leache, A. D. 2006. Direct and indirect effects of environmental temperature on the evolution of reproductive strategies: an information-theoretic approach. *American Naturalist*, **168**, E123–E135.

Arnold, S. J. 1992. Constraints on phenotypic evolution. *American Naturalist*, **140**, S85–S107.

Arnold, S. J. & Phillips, P. C. 1999. Hierarchical comparison of genetic variance–covariance matrices. II. Coastal-inland divergence in the garter snake, *Thamnophis elegans*. *Evolution*, **53**, 1516–1527.

Barrett, P. 2007. Structural equation modelling: adjusting model fit. *Personality and Individual Differences*, **42**, 815–824.

Bell, A. M. 2005. Behavioral differences between individuals and two populations of stickleback (*Gasterosteus aculeatus*). *Journal of Evolutionary Biology*, **18**, 464–473.

Bell, A. M. 2007. Future directions in behavioural syndromes research. *Proceedings of the Royal Society B*, **274**, 755–761.

Bell, A. M. & Sih, A. 2007. Exposure to predation generates personality in three-spined sticklebacks (*Gasterosteus aculeatus*). *Ecology Letters*, **10**, 828–834.

Bell, A. M. & Stamps, J. A. 2004. Development of behavioural differences between individuals and populations of sticklebacks, *Gasterosteus aculeatus*. *Animal Behaviour*, **68**, 1339–1348.

Bentler, P. M. & Yuan, K. H. 1999. Structural equation modeling with small samples: test statistics. *Multivariate Behavioral Research*, **34**, 181–197.

Biro, P. A. & Stamps, J. A. 2008. Are animal personality traits linked to life-history productivity? *Trends in Ecology & Evolution*, **23**, 361–368.

Blows, M. W. 2007. Complexity for complexity's sake? *Journal of Evolutionary Biology*, **20**, 39–44.

Brodie, E. D. & McGlothlin, J. W. 2007. A cautionary tale of two matrices: the duality of multivariate abstraction. *Journal of Evolutionary Biology*, **20**, 9–14.

Brown, B. A. 2006. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.
Brydges, N. M., Colegrave, N., Heathcote, R. J. P. & Braithwaite, V. A. 2008. Habitat stability and predation pressure affect temperament behaviours in populations of three-spined sticklebacks. *Journal of Animal Ecology*, **77**, 229–235.
Burnham, K. P. & Anderson, D. R. 2002. *Model Selection and Multimodel Inferences: a Practical Information-Theoretic Approach*. New York: Springer-Verlag.
Cade, W. H. & Cade, E. S. 1992. Male mating success, calling and searching behaviour at high and low densities in the field cricket, *Gryllus integer*. *Animal Behaviour*, **43**, 49–56.
Clark, A. B. & Ehlinger, T. J. 1987. Pattern and adaptation in individual behavioral differences. In: *Perspectives in Ethology* (Ed. by P. P. G. Bateson & P. H. Klopfer), pp. 1–47. New York: Plenum.
Coleman, K. & Wilson, D. S. 1998. Shyness and boldness in pumpkinseed sunfish: individual differences are context specific. *Animal Behaviour*, **56**, 927–936.
Dietz, E. J. 1983. Permutation tests for association between 2 distance matrices. *Systematic Biology*, **32**, 21–26.
Dingemanse, N. J. & Réale, D. 2005. Natural selection and animal personality. *Behaviour*, **142**, 1165–1190.
Dingemanse, N. J., Wright, J., Kazem, A. J. N., Thomas, D. K., Hickling, R. & Dawnay, N. 2007. Behavioural syndromes differ predictably between 12 populations of stickleback. *Journal of Animal Ecology*, **76**, 1128–1138.
Dingemanse, N. J., van der Plas, F., Wright, J., Réale, D., Schrama, M., Roff, D. A., van der Zee, E. & Barber, I. 2009. Individual experience and evolutionary history of predation affect expression of heritable variation in fish personality and morphology. *Proceedings of the Royal Society B*, **276**, 1285–1293.
Dingemanse, N. J., Kazem, A. J. N., Réale, D. & Wright, J. In press. Behavioural reaction norms: where animal personality meets individual plasticity. *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2009.07.013.
Dochtermann, N. A. & Jenkins, S. H. 2007. Behavioural syndromes in Merriam's kangaroo rats (*Dipodomys merriami*): a test of competing hypotheses. *Proceedings of the Royal Society B*, **274**, 2343–2349.
Dochtermann, N. A. & Jenkins, S. H. In press. Developing and evaluating multiple hypotheses in behavioral ecology. *Behavioral Ecology and Sociobiology*.
Garamszegi, L. Z. 2006. Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behavioral Ecology*, **17**, 682–687.
Garcia, L. V. 2004. Escaping the Bonferroni iron claw in ecological studies. *Oikos*, **105**, 657–663.
Gosling, S. D. 2001. From mice to men: what can we learn about personality from animal research? *Psychological Bulletin*, **127**, 45–86.
Gosling, S. D. & John, O. P. 1999. Personality dimensions in nonhuman animals: a cross-species review. *Current Directions in Psychological Science*, **8**, 69–75.
Grace, J. B. 2006. *Structural Equation Modeling and Natural Systems*. Cambridge: Cambridge University Press.
Grace, J. B. & Jutila, H. 1999. The relationship between species density and community biomass in grazed and ungrazed coastal meadows. *Oikos*, **85**, 398–408.
Haughton, D. M. A., Oud, J. H. L. & Jansen, R. A. R. G. 1997. Information and other criteria in structural equation model selection. *Communications in Statistics: Simulation and Computation*, **26**, 1477–1516.
Herzog, W. & Boomsma, A. 2009. Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling: Multidisciplinary Journal*, **16**, 1–27.
Huntingford, F. A. 1976. The relationship between anti-predator behaviour and aggression among conspecifics in the three-spined stickleback, *Gasterosteus aculeatus*. *Animal Behaviour*, **24**, 245–260.
Kline, R. B. 2005. *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.
Kohn, L. A. P. & Atchley, W. R. 1988. How similar are genetic correlation structures: data from mice and rats. *Evolution*, **42**, 467–481.
Lofsvold, D. 1986. Quantitative genetics of morphological differentiation in *Peromyscus*. 1. Tests of the homogeneity of genetic covariance structure among species and subspecies. *Evolution*, **40**, 559–573.
Martin, J. G. A. & Réale, D. 2008. Temperament, risk assessment and habituation to novelty in eastern chipmunks, *Tamias striatus*. *Animal Behaviour*, **75**, 309–318.
Mather, J. A. & Anderson, R. C. 1993. Personalities of octopuses (*Octopus rubescens*). *Journal of Comparative Psychology*, **107**, 336–340.
Mettke-Hofmann, C. 2007. Context-specific neophilia and its consequences for innovations. *Behavioral and Brain Sciences*, **30**, 419–420.
Mettke-Hofmann, C., Winkler, H. & Leisler, B. 2002. The significance of ecological factors for exploration and neophobia in parrots. *Ethology*, **108**, 249–272.
Mettke-Hofmann, C., Wink, M., Winkler, H. & Leisler, B. 2005. Exploration of environmental changes relates to lifestyle. *Behavioral Ecology*, **16**, 247–254.
Moretz, J. A., Martins, E. P. & Robison, B. D. 2007. Behavioral syndromes and the evolution of correlated behavior in zebrafish. *Behavioral Ecology*, **18**, 556–562.
Petraitis, P. S., Dunham, A. E. & Niewiarowski, P. H. 1996. Inferring multiple causality: the limitations of path analysis. *Functional Ecology*, **10**, 421–431.
Phillips, P. C. & Arnold, S. J. 1999. Hierarchical comparison of genetic variance–covariance matrices. I. Using the Flury hierarchy. *Evolution*, **53**, 1506–1515.
Quinn, G. P. & Keoug, M. J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.
Réale, D., Reader, S. M., Sol, D., McDougall, P. & Dingemanse, N. J. 2007. Integrating temperament in ecology and evolutionary biology. *Biological Reviews*, **82**, 291–318.
Reddon, J. R. & Jackson, D. N. 1984. A note on testing the sphericity hypothesis with Bartlett test. *Multivariate Experimental Clinical Research*, **7**, 49–52.
Rencher, A. C. 2002. *Methods of Multivariate Analysis*. New York: J. Wiley.
Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology*, **86**, 2805–2814.
Riechert, S. E. & Hedrick, A. V. 1993. A test of correlations among fitness-related behavioural traits in the spider, *Agelenopsis aperta* (Aranea, Agelinadae). *Animal Behaviour*, **46**, 669–675.
Roff, D. A. 2002. Comparing G matrices: a MANOVA approach. *Evolution*, **56**, 1286–1291.
Roff, D. A., Mousseau, T. A. & Howard, D. J. 1999. Variation in genetic architecture of calling song among populations of *Allonemobius socius*, *A. fasciatus*, and a hybrid population: drift or selection? *Evolution*, **53**, 216–224.
Sgro, C. M. & Hoffmann, A. A. 2004. Genetic correlations, tradeoffs and environmental variation. *Heredity*, **93**, 241–248.
Shaw, R. G. 1991. The comparison of quantitative genetic parameters between populations. *Evolution*, **45**, 143–151.
Sih, A. & Bell, A. M. 2008. Insights for behavioral ecology from behavioral syndromes. *Advances in the Study of Behavior*, **38**, 227–281.
Sih, A., Bell, A. M., Johnson, J. C. & Ziemba, R. E. 2004. Behavioural syndromes: an integrative overview. *Quarterly Review of Biology*, **79**, 241–277.
Sinn, D. L., Gosling, S. D. & Moltschaniwskyj, N. A. 2008. Development of shy/bold behaviour in squid: context-specific phenotypes associated with developmental plasticity. *Animal Behaviour*, **75**, 433–442.
Stamps, J., McElreath, R. & Eason, P. 2005. Alternative models of conspecific attraction in flies and crabs. *Behavioral Ecology*, **16**, 974–980.
Steppan, S. J., Phillips, P. C. & Houle, D. 2002. Comparative quantitative genetics: evolution of the G matrix. *Trends in Ecology & Evolution*, **17**, 320–327.
Tabachnick, B. G. & Fidell, L. S. 2001. *Using Multivariate Statistics*. Boston, Massachusetts: Allyn & Bacon.
Violato, C. & Hecker, K. G. 2007. How to use structural equation modeling in medical education research: a brief guide. *Teaching and Learning in Medicine*, **19**, 362–371.
Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.
Zar, J. H. 1999. *Biostatistical Analyses*. Upper Saddle River, New Jersey: Prentice-Hall.

APPENDIX

Distinguishing between Models of Syndrome Structure at Applicable Sample Sizes

Structural equation modelling (SEM) is a statistical approach that allows the assessment of multiple interacting factors and potential inferences regarding causation (Grace 2006). This approach was initially intended to allow the determination of how biotic and abiotic factors interact (Wright 1921) and can be used to assess a variety of ecological and evolutionary questions. SEM estimates the strength of the relationships between observed causal factors, allows the estimation of unobserved causal factors and allows the estimation of how much these causal factors vary. SEM also enables the assessment of complex and nonlinear interactions, comparisons between populations and numerous other applications that are not readily implemented with the usual statistical methods of ecologists (Grace 2006). However, the applicability of the approach to ecological research has been criticized for a variety of reasons, including sample size requirements (e.g. Petraitis et al. 1996). It is the issue of sample size that we detail here.

Like other multivariate approaches, SEM generally requires large sample sizes (Brown 2006). Barrett (2007) suggested that SEM should never be conducted without sample sizes of at least 200. Sample size concerns for SEM use generally fall into one of two categories. First, assessment of model fit is based on the discrepancy between a structural model and the data. This discrepancy is often evaluated against a chi-square distribution to determine whether the model departs significantly from the data. Because these discrepancy estimates may not conform to chi-square distributions at smaller sample sizes, large samples are needed (Bentler & Yuan 1999; Kline 2005; Barrett 2007). Second, because the discrepancies are estimated based on the overall variance–covariance matrix and the degrees of freedom for their testing is based on model structure, tests of model fit often suffer from a profound lack of power.

However, whether available estimates of sample sizes are applicable to ecological research is worth questioning. For example, the influential simulations of Bentler & Yuan (1999) suggested that sample sizes below 100–200 almost always suffer from unacceptable type I or type II errors, limiting the applicability of SEM in ecological research since these sample sizes may not be logistically feasible. Unfortunately, this simulation was conducted with structural models in which over 30 parameters were estimated (Bentler & Yuan 1999). In contrast, structural equation models used in ecological research often estimate far fewer parameters. For example, when using SEM to ask how temperatures affect reproductive strategies in eastern fence lizards, *Sceloporus undulatus*, Angilletta et al. (2006) evaluated models in which only 11–13 parameters were estimated. Dochtermann & Jenkins (2007) conducted confirmatory factor analysis using SEM methods to determine how behaviours covary and their models required the estimation of even fewer (eight) parameters. Similarly, the worked example discussed in this paper is based on only 10 parameters. Along with evaluating much simpler models, Angilletta et al. (2006) and Dochtermann & Jenkins (2007) and the stickleback example discussed above also used much smaller sample sizes; 19, 19 and 42, respectively. These uses of SEM clearly violate the general recommendations of sample size requirements for the use of SEM.

In addition to differing in the complexity of models for which required sample size recommendations have been generated, these ecological examples differ as well in that evidence for particular structural models was not evaluated in an absolute sense (Angilletta et al. 2006; Dochtermann & Jenkins 2007; see also above). Instead, each of these three examples used a model comparison approach based on the Akaike information criterion (AIC). Thus these three examples determined the relative differences in how far proposed models diverged from the available data (Burnham & Anderson 2002). Unfortunately, the performance of AIC values and other information criteria has not been evaluated at sample sizes relevant to any of these three examples.

Haughton et al. (1997) evaluated the performance of several information criteria, including AIC and the Bayes/Schwaz information criterion (BIC), based on extensive simulations. These results suggested that information criteria generally performed well in distinguishing between models, although AIC use led to some general overfitting. The applicability of these results to ecological research is, however, questionable because of the sample sizes used (100–6000; Haughton et al. 1997).

Here, we extend the general simulation framework of Haughton et al. (1997) to sample sizes that may be more applicable to both to ecological researchers, and to our worked example discussed above.

METHODS

Following Haughton et al. (1997), we generated multivariate normal data based on known variances and covariances. We used two (co)variance matrices to do so:

$$\text{(A)}\quad \begin{pmatrix} 1 & 0.2 & 0.2 & 0.2 & 0 \\ 0.2 & 1 & 0.2 & 0.2 & 0 \\ 0.2 & 0.2 & 1 & 0.2 & 0 \\ 0.2 & 0.2 & 0.2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{(B)}\quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The first of these (co)variance matrices, 'A', represents data that could be modelled with four variables covarying because of underlying connections to a latent variable while the fifth variable varies independently. The second (co)variance matrix, 'B', represents data that are all causally independent of one another. This matrix was used to determine whether the information criterion leads to overfitting to a degree where causal inferences would be drawn despite variables being independent of one another. These matrices were each used to generate 20 data sets of $N = 42$ using MATLAB (MathWorks, Natck, MA, U.S.A.). This sample size was used to mimic the conditions of the worked example given above. Each of the data sets was analysed using the program AMOS 7.0. AMOS was also used in each of the three ecological examples discussed earlier.

The performances of AIC and BIC were determined based on their estimation of the correspondence between the simulated data and each of three models. These models correspond to model structures recently advocated for use in both evolutionary ecological questions regarding behavioural covariance (see above), as well as for use in comparing trait distributions between populations based on either phenotypic or genotypic (co)variance matrices (Dochtermann & Jenkins, in press). The three models considered were: (1) all five variables covary because of the effects of an underlying latent variable; (2) the four variables with an average correlation of 0.2 covary because of the effects of an underlying latent variable while the fifth varies independently; (3) all five variables vary independently. Model 1 represents an overfitted model for data generated with (co)variance matrix 'A' and 'B'. Model 2 corresponds to the underlying structure used to generate the data from (co)variance matrix 'A' but is overfitted for data generated with (co)variance matrix 'B'. Model 3 is substantially underfitted for data generated by (co)variance matrix 'A' but corresponds to (co)variance matrix B. These models are described graphically in Fig. A1.

AMOS was used to estimate the discrepancy ($\hat{C}$) between a proposed model (Fig. A1a–c) and the simulated data. $\hat{C}$ was then used to calculate a model's AIC and BIC for each data set. AIC and BIC were calculated as:

$$\text{AIC} = \hat{C} + 2k \quad \text{and} \quad \text{BIC} = \hat{C} + k\ln(N)$$

where $\hat{C}$ is a model's discrepancy, k is the number of parameters estimated for a structural model and N is the sample size. Comparisons between models within a data set were then made based on standard criteria; if two models differed by fewer than 2 AIC points, we considered them to correspond to the data equally well (Burnham & Anderson 2002; Richards 2005). We used the same criterion for evaluating models based on BIC. More specifically, if the model of variable independence (Fig. A1c) did not have an AIC or BIC value more than 2 points greater than that of either other model then we considered the hypothesis that variables were independent of one another to be equally well supported. This model functions effectively as a null expectation and corresponds to a structural independence model.

We used both of these two (co)variance matrices because the performance of AIC and BIC for model comparison could then be used to estimate discriminatory ability comparable to both type I and type II error rates. If model 3 was supported as well as either model 1 or 2 for the data generated from (co)variance matrix 'A', this would be comparable to a type II error. If model 3 were not supported at least as well as either model 1 or 2 for data generated from (co)variance matrix 'B', this would be comparable to a type I error. Since model 2, which corresponds to (co)variance matrix 'A', is nested within model 1, if either information criterion ranked model 1 above model 2 this would be considered overfitting. Overfitting for data generated with (co)variance matrix 'A' is not considered to represent poor performance by a particular criterion because both models contain the generating structure.

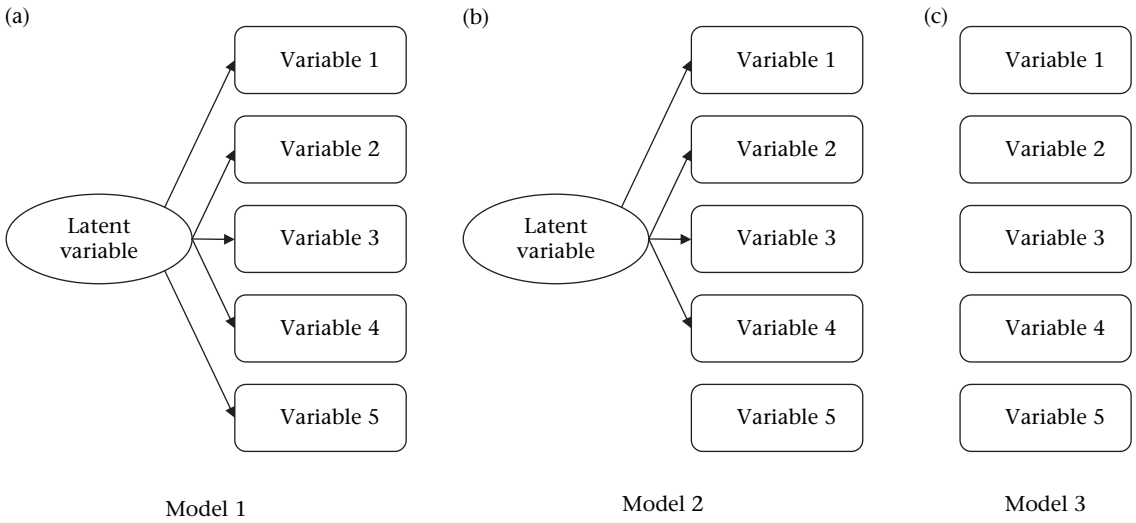

Figure A1. Graphical representation of the three models used to evaluate discriminatory power of information criteria and model selection approaches. Ovals correspond to unmeasured variables (latent variables) and rectangles correspond to observed variables. Latent variables influence the observed variables if connected by arrows. Model 1 corresponds to covariance between all five variables because they are connected by an underlying latent variable. Model 2 is the same as model 1 except variable 5 is allowed to vary independently. Model 3 represents a situation where the observed variables are independent of one another.

RESULTS AND DISCUSSION

For data generated from a distribution in which variables did covary, AIC values tended to rank models properly (Table A1). In all cases, model 3 was ranked below models 1 or 2 (Table A1). Model 3 typically (75%) differed from either 1 or 2 by more than 2 AIC values (Table A1). In contrast, BIC values only ranked model 3 more than 2 BIC values or higher than models 1 and 2 40% of the time (Table A1). This demonstrates that AIC values led to a ranking of models with an acceptable (0.25) rate of failure to exclude an incorrect null. If this were considered comparable to a type II error rate then with these data AIC values had a power of 0.75 while the power of BIC values to rank SEM models properly was only 0.40.

For data generated from a distribution in which variables did not covary, both AIC and BIC values ranked models properly (Table A2). In all cases, model 3 was ranked higher or within 2 points of models 1 and 2. This demonstrates that AIC and BIC values led to a ranking of models with a more than acceptable, 0.00, rate of excluding a null. If this were considered comparable to a type I error rate then with these data AIC and BIC values had an error rate (α) of 0 which would make their use more conservative in this regard than χ^2 values. It is surprising that the increased 'power' afforded by the use

Table A1
AIC and BIC values for each of three models for 20 simulated data sets generated from (co)variance matrix A

Data set	AIC			BIC		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
1	25.601	24.016	33.11	42.977	39.655	41.799
2	25.518	26.54	**26.93**	42.894	42.179	35.619
3	23.143	21.933	**21.44**	40.519	37.572	30.128
4	23.191	22.112	37.595	40.567	37.751	46.284
5	27.79	29.446	39.496	45.167	45.085	48.634
6	21.938	21.929	29.256	39.315	37.568	37.944
7	25.384	24.321	27.969	42.76	39.96	36.658
8	24.606	22.986	28.094	41.982	38.625	36.782
9	25.934	23.964	35.436	42.771	39.603	44.125
10	30.146	28.221	30.231	47.522	43.86	38.919
11	25.02	23.735	36.891	42.397	39.374	45.579
12	26.913	26.797	47.683	44.29	42.436	56.397
13	24.561	22.575	41.489	41.938	38.214	50.177
14	21.744	20.555	**20.45**	39.121	36.194	29.139
15	26.216	26.193	33.244	43.593	41.832	41.932
16	23.847	28.361	41.479	41.223	44	50.168
17	25.621	25.991	28.39	42.998	41.63	37.079
18	28.697	27.488	**28.187**	46.073	43.127	36.875
19	22.632	20.671	**17.961**	40.009	36.31	26.649
20	22.124	20.739	36.378	39.501	36.378	26.729

This (co)variance matrix is consistent with model 2 (see text). AIC rankings generally lead to the proper exclusion of model 3 while BIC rankings did not. In five instances (printed in bold), AIC-based ranking included model 3, the model of variable independence, within 2 AIC values of the other two models. These five occurrences can be considered analogous to type II errors relative to the generating distribution. In the context of type II errors, this would suggest that AIC-based ranking has a power of 0.75.

Table A2
AIC and BIC values for each of three models for 20 simulated data sets generated from (co)variance matrix 'B'

Data set	AIC			BIC		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
1	22.287	20.31	17.825	39.664	35.949	26.513
2	23.955	23.42	25.099	41.331	39.06	33.788
3	24.514	22.721	18.964	41.89	38.36	27.652
4	23.126	25.23	23.955	40.502	40.869	32.643
5	21.303	27.636	20.747	38.679	43.275	29.435
6	21.995	23.448	20.731	39.372	39.087	29.42
7	24.165	22.457	21.307	41.542	38.096	29.996
8	21.355	24.505	17.619	38.732	40.144	26.308
9	21.021	19.844	13.73	38.398	35.483	22.418
10	21.641	22.294	16.213	39.018	37.933	24.901
11	20.596	18.711	12.919	37.972	34.35	21.608
12	24.993	25.161	26.658	42.37	40.8	35.346
13	21.633	21.609	19.037	39.01	37.248	27.725
14	20.504	21.177	13.79	37.881	36.816	22.479
15	24.209	26.728	27.073	41.586	42.367	35.761
16	20.934	19.369	13.165	38.31	35.008	21.853
17	27.372	25.344	24.009	44.749	40.983	32.697
18	21.093	22.238	15.145	38.47	37.877	23.833
19	21.84	26.152	19.769	39.216	41.791	28.457
20	23.962	22.934	19.126	41.339	38.573	27.815

This (co)variance matrix is consistent with model 3 (see text). AIC and BIC rankings lead to model 3 being consistently ranked as high or higher than either of the other models. Thus, both AIC- and BIC-based rankings properly controlled for the improper exclusion of null expectations. Considered in the context of null hypothesis testing, these results are analogous to a type I error rate less than the standard expectation of 0.05.

of AIC values did not result in a concordant increase in the rate of error comparable to type I errors.

In all cases, model 1 received at least as much support within data generated by covariance matrix A as model 2. This suggests that with AIC, if a model corresponding to the generating model is nested within another there may be some tendency towards overfitting. This is consistent with Haughton et al. (1997), who found that overfitting with AIC occurred only when the generating model was nested within another model.

GENERAL DISCUSSION

These results suggest that at sample sizes much smaller than those generally considered acceptable for the application of SEM, AIC use is an acceptable method for evaluating relative model fits. In contrast, BIC use is unacceptably conservative. These results also suggest that while at small sample sizes discrepancy estimates may not conform to chi-square distributions, they are not biased in their evaluation of relative fit. Based on these results it is reasonable to suggest that SEM analyses can be conducted with sample sizes much smaller than those typically recommended, if discrepancy estimates are used to evaluate relative fit rather than absolute fit. Thus we recommend that researchers interested in using SEM within an ecological framework but that are concerned about sample size requirements consider using an approach based on the evaluation of a priori models ranked based on AIC use (see recommendations given above). Despite the potential for some overfitting with AIC use, BIC values are not recommended owing to their improper power to exclude null models properly from consideration.