Behavioral Ecology

The official journal of the
ISBE
International Society for Behavioral Ecology

Behavioral Ecology (2013), 24(5), 1068–1080. doi:10.1093/beheco/art033

Original Article

A meta-analysis of correlated behaviors with implications for behavioral syndromes: relationships between particular behavioral traits

László Zsolt Garamszegi[a], Gábor Markó,[b,c] and Gábor Herczeg[b,d]
[a]Department of Evolutionary Ecology, Estación Biológica de Doñana–CSIC, c/Americo Vespucio, s/n, 41092 Seville, Spain, [b]Behavioural Ecology Group, Department of Systematic Zoology and Ecology, Eötvös Loránd University, Pázmány Péter sétány 1/c, H-1117 Budapest, Hungary, [c]Department of Plant Pathology, Corvinus University of Budapest, Ménesi út 44, H-1118 Budapest, Hungary, and [d]Ecological Genetics Research Unit, Department of Biosciences, University of Helsinki, PO Box 65, FI-00014 Helsinki, Finland

Received 30 July 2012; revised 20 March 2013; accepted 21 March 2013; Advance Access publication 5 June 2013.

Behavioral syndromes predict that individuals display behaviors consistently across different ecological situations, resulting in correlations among functionally different individual-specific behaviors (e.g., activity, exploration, aggression, and risk taking). Such consistencies can arise because of the common innate government of traits (i.e., temperament). However, different behaviors can be mediated by different selection regimes and/or measured with different errors. Furthermore, contextual overlap among traits may also vary. These possibilities can cause dissimilarities in the pair-wise relationship between particular traits. To determine the relationships among the most studied behaviors, we performed a modern meta-analysis, in which we assessed the strength of correlations in each possible combination of traits. Relying on data from 81 scientific papers, we found that the correlations among behaviors were generally weak and that they varied in magnitude across comparisons (e.g., novel environment exploration and activity: $r = 0.345$; novel object exploration and activity: $r = 0.074$). The partial correlations among traits revealed that certain relationships (e.g., novel environment exploration/activity and the novel object exploration/risk taking) were independent of the covariation with other traits, whereas certain relationships (e.g., aggression/novel environment exploration) consistently weakened after controlling for covariance. Some relationships were affected by contextual overlap: the effect sizes were systematically higher when the behaviors were assayed in the same experimental compartment (e.g., same test room or aquarium). Different correlations are unlikely to emerge due to differences in repeatabilities that are associated with the measurement of different traits, as we found that averaged repeatabilities vary around the same intermediate magnitude for each behavior. We suggest that the most commonly assessed behavioral traits do not necessarily form equally independent domains.

Key words: **behavioral consistency, behavioral type, coping style, meta-regression, personality, phylogenetic meta-analysis, syndrome deviation, temperament.**

INTRODUCTION

One of the most exciting discoveries of current research in evolutionary ecology of behavior is that apparently different behaviors expressed by individual animals are not necessarily independent of each other, but often correlate and form a behavioral syndrome (Sih, Bell, Johnson, & Ziemba 2004; Sih and Bell 2008; Sih et al. 2012). Such a linkage can have important implications for the optimization of life-history trade-offs (Wolf et al. 2007), for the reliability of sexual signals (Garamszegi et al. 2008; Logue et al. 2009; Mateos-González and Senar 2012), and for the interaction with parasites and predators (Quinn and Cresswell 2005; Huntingford and Coyle 2007; Barber and Dingemanse 2010; Coats et al. 2010). Several candidate hypotheses were put forward to explain why and how apparently different behaviors correlate (Bell

Address correspondence to L.Z. Garamszegi. E-mail: laszlo.garamszegi@ebd.csic.es.

2005; Wolf and Weissing 2010), but so far none of these has been proven to provide a universal account for the widespread existence of behavioral syndromes in animals. This is because correlations are meaningful when comparing groups of individuals (i.e., population level) only, and for a robust evolutionary interpretation, comparisons among populations or species with different syndrome structure would be needed. However, most studies rarely go beyond reporting the existence or absence of behavioral syndromes within a single population (Herczeg and Garamszegi 2012).

Although researchers studying syndromes in different species measure various behavioral traits, these traits can be categorized into 5 main axes that correspond to different ecological situations (Réale et al. 2007; Conrad et al. 2011). The "big five" framework (or five-factor model) resembles the approach used in human psychology, in which openness, conscientiousness, extraversion, agreeableness, and neuroticism are treated as the main dimensions of human personality (McCrae and John 1992; Carver and Connor-Smith 2010). However, in evolutionary ecology, domains are defined according to the prevailing environmental challenge under which the behavior in question is expressed (Réale et al. 2007): activity (general activity under no environmental or social challenges), exploration (in the presence of novelty), boldness or risk-taking (in the presence of predator or stress stimuli), aggression (in the presence of intraspecific competitor), and sociability (in any nonaggressive social contact). These 5 axes are often labeled as temperament traits or personality traits (Réale et al. 2007), but we simply refer to them as "behavioral traits" to circumvent confusion between terminologies.

If the studied domains are functionally independent, a correlation between them can have evolutionary significance. One possibility is that different selection forces shape each functionally distinct behavior, and the observed correlation among them in a population is the net effect of the locally acting selection processes (Wilson 1998; Bell 2005; Bell and Sih 2007). In a particular environment, for example, high predation rate combined with limited mating opportunities can favor risk-averse but aggressive individuals, resulting in a negative correlation between risk taking and aggression but without consequence for activity. Conversely, it is also plausible that different behaviors are expressions of the same background "trait," thus not particular behaviors, but the individual background "trait" is the subject of selection. Consistency both within and among behaviors can be caused by, for instance, temperament-like drives that affect multiple aspects of behavioral phenotypes during life (Buss et al. 1987; Clark and Wilson 1999; Gosling 2001). Furthermore, mechanisms based on stable states and state-dependent behaviors also require that the correlation among traits is mediated by the same individual-specific attribute, such as age, size, or condition (Wolf and Weissing 2010). The effect of these common underlying factors can be manifested in the systematic display of shy or bold behaviors in different ecological situations, leading to behavioral correlations, as aggressive individuals will necessarily be the risk-takers and explorative ones. If the preservation of both shy and bold phenotypes is favored in the long run (e.g., due to the unpredictably changing environment), generally positive correlations among different behaviors will be observed in the population.

The identification of the "big five" domains from the evolutionary perspective seems convenient, because it offers a standardized framework to study behavioral syndromes via well-defined and distinct categories of traits (Réale et al. 2007; Conrad et al. 2011). This helps to avoid the identification of false syndromes formed by behaviors from seemingly different domains, which are actually different measures of the same behavior within a single domain. However, this approach inherently assumes that behaviors are at least contextually independent, although this is not necessarily the case. Observational data can easily yield spurious correlations among behaviors, when the detected associations are actually caused by an unknown/uncontrolled "third" factor without having any consequence for a syndrome. For example, if one measures exploration or aggression in a particular test chamber without proper acclimation, intertrait correlations may appear due to individual consistency in stress response to an artificial environment (Maier et al. 1988; Budaev 1997). Similarly, if behaviors are assessed within the same territory in the natural habitat, detected correlations among them may reflect how individuals perceive the value of the territory, as males defending resources perceived as good quality are likely to fight for it aggressively and take high risk in the presence of predators (Garamszegi et al. 2009). Such obvious problems may be eliminated by a careful study design, but certain bias due to different levels of contextual overlap will likely remain inevitable. This is because the physiological states and life cycles of animals continuously change and thus cannot be sorted into independent categories (e.g., Wingfield and Farner 1993). Therefore, even if all measured individual behaviors were scored repeatedly in different test chambers and in different days or weeks, they may still correspond to partially similar hormonal and/or breeding conditions.

From the literature, it seems that the evidence for the relationship among behavioral traits is mixed. For example, in an avian model, namely the great tit *Parus major*, novel environment exploration was related to male responses to a territorial intrusion stimulus (Verbeek et al. 1996; Amy et al. 2010), whereas in another study, aggression toward a caged or free intruder was not significantly associated with activity in the novel environment (Carere et al. 2005). Inconsistencies were also found even in the same study. In the great tit, during wintering, the relationship between dominance rank and exploration score was negative in juvenile males, but it was positive in adults (Dingemanse and de Goede 2004). In the pied flycatcher, *Ficedula hypoleuca*, there was a positive and significant relationship between latency to approach a novel object and latency to resume activity after a predator attack, but the same tendency was not observed between estimates of novel environment exploration and risk-taking (Ruuskanen and Laaksonen 2010). In another well-studied model in behavioral syndrome research, the three-spined stickleback *Gasterosteus aculeatus*, strong correlations among exploration, boldness, and aggression were detected in high-predation populations but not in low-predation populations (Bell 2005; Dingemanse et al. 2007), and these patterns were similar in laboratory-reared fish (Bell and Stamps 2004). In another study, Brydges et al. (2008) found almost no sign of exploration–boldness syndrome using several ecologically divergent stickleback populations.

Such extensive variation in the strength and even the direction of correlations allows investigating evolutionary questions about behavioral syndromes at the above-individual level, as differences in study outcomes likely result from differences in the studies' ecological (and methodological) backgrounds. When inconsistent patterns emerge in the literature, meta-analyses can be applied as a statistical basis for theoretical generalizations (Arnqvist and Wooster 1995; Lajeunesse and Forbes 2003; Adams 2008; Borenstein et al. 2009). In this kind of quantitative review, study results are transformed into a common metric, the so-called "effect size," that can be compared across studies. Each study is weighted by its precision (sample size or confidence interval [CI]) and further approaches are available that can correct other sources of bias such as phylogenetic inertia or publication bias. The key benefit of meta-analysis is that it also allows testing of heterogeneities in effect sizes across studies, and one can test the effect of different predictors that can potentially mediate the variation of effect sizes among studies, to seek factors that explain variation in the degree of association among particular behavioral traits.

In a previous meta-analysis (Garamszegi, Markó, et al. 2012), we tested for general patterns that mediate the degree of correlation among behaviors of any kind. In that study, we investigated how issues related to taxonomy, phylogeny, publication bias, repeatability, and study/sampling design can affect the overall relationship among behaviors irrespective of their categorization. In this paper, our main aim was to conduct a modern meta-analysis to compare the absolute magnitude and direction of the relationship between different particular pairs of traits. Here, by differentiating among behavioral domains, we target questions about the common organization of traits and focus on how contextual nonindependence and differences in repeatability of behaviors can drive differences in main effect sizes calculated for particular domains. If behavioral syndromes are composed of the homogeneous combination of functionally and contextually independent behaviors because they are different manifestations of the same background trait (e.g., temperament), and if they can be measured with the same error, we would expect all trait pairs to show similar relationships in the same direction. However, if different correlations 1) have different adaptive value, 2) have different genetic or hormonal constraints, 3) are shaped by different methodological constraints, or 4) are loaded with different measurement errors, we would expect variation in the strength and perhaps even direction of correlations. Furthermore, if the established category domains incorporate nonindependent behaviors, the correlations among particular traits will vary depending on the similarity in the context among domains. Accordingly, we examined the mean correlation among 4 main behavioral domains and tested whether spatial overlap (tests were performed in the same test arena, territory, etc.) and temporal overlap (time elapsed between subsequent behavioral tests) can explain some proportion of variation in strength and direction of behavioral correlations. Finally, if correlations vary among different pair-wise comparisons of traits because behaviors can be assessed with different errors, given that $r_{(\text{observed})\text{A,B}} = r_{\text{A,B}} \times \sqrt{\text{repeatability}_{\text{A}} \times \text{repeatability}_{\text{B}}}$ for behaviors A and B and based on Spearman (1904), the observed correlation $r_{(\text{observed})\text{A,B}}$ should be weaker among traits that are assessed with lower repeatability.

METHODS

Data set

For the details of the literature search, please see the electronic supplementary material of Garamszegi, Markó, et al. (2012, http://link.springer.com/content/esm/art:10.1007/s10682-012-9589-8/file/MediaObjects/10682_2012_9589_MOESM1_ESM.doc). Briefly, based on the combination of keyword and cross-reference searches, we collected published scientific papers that potentially hold information on the correlation among the behavioral traits depicting activity, aggression, exploration, and risk-taking (we did not deal with sociability, as only a handful of studies provided its correlations). To test our hypotheses about the nonindependence of domains in a meta-analysis framework, we summarized data from 81 scientific papers (published between 1976 and 31 August 2011) reporting behavioral correlations in 58 species (Supplementary Table S1). Due to the complexity of terminologies used in the corresponding literature, we cannot exclude the possibility that we failed to find some relevant papers and so must infer that our large sample size holds sufficient power to detect general tendencies. We also assumed that the papers we had missed in our search represent a random sample with respect to the aims of our meta-analysis such that the available data are not biased. In any case, to combat the potential problems that might be caused by undiscovered effect sizes as well as to incorporate future studies (see Garamszegi, Nunn, et al. 2012), we also upload our data to a dynamically developing repository of meta-analysis data (http://evolutionary-meta-analysis.net/).

Because the behavioral syndrome (and animal personality) literature does not strictly follow a consistent terminology and approach for the definition of variables, we were constrained to apply some consensus for the inclusion of different behavioral variables based on our predefined judgments, in which we tried to accommodate the suggestions of Réale et al. (2007). Accordingly, we labeled "activity" those variables that could be used to describe the general intensity of movements in a familiar environment without any social and environmental challenge. In this sense, we followed a liberal definition and included all variables in this category that were measured in any context including wild and captive conditions. However, we also worked along a stricter framework, in which we only considered activity traits that were measured in the control situation of the experimental challenge situation (e.g., movement activity in the familiar environment when tested against movement activity in an unfamiliar environment, or behavioral intensity prior to the presentation of novel object, conspecific, or predator challenge). "Exploration" is usually estimated as the intensity of movements when exposing individuals to a certain novel stimulus but is often measured in 2 different setups. Either it is assayed in a situation when an animal is placed in a completely novel environment or the behavior is scored in a familiar environment but in the presence of a novel object. We distinguished between "novel environment exploration" (i.e., variables that reflected the intensity of movements or the exploitation of space in the former situation) and "novel object exploration" (i.e., variables that reflected the intensity of movements in the presence of a new object or the approach time/distance toward the novelty). Based on our broad-sense definitions, we included variables irrespective of whether they were measured in captivity or in the wild or whether the animals were motivated with some reward such as food or access to a mate. Nevertheless, we also developed a stricter consensus, in which we only included exploration traits if they originated from a real novel environment challenge experiment (i.e., the animals were placed in an unfamiliar test compartment) and they were estimated without the use of a reward in case of testing novel object response. We derived correlations for "risk-taking" as a measure of boldness if the focal variable mirrored the intensity of movements or behavioral activity in or after a perceived life-threatening situation. This open-minded definition extends to any major stress factor, such as predator (including human) or startle stimulus and involves estimates of latency of resuming normal activity after the challenge or estimates of actual contact rate (e.g., inspection) with or distance from stress stimuli. We also created a less permissive agenda for the study of risk taking, in which we only considered variables that corresponded to a predator stimulus. Finally, we gathered information on "aggression" when the measured trait depicted the intensity of an antagonistic behavior in a social conflict (such as a conspecific individual of the same sex or a mirror image) or the approach time/distance toward the opponent. This category also involved estimates of social ranks among individuals reflecting the consequence of pair-wise aggressive interactions in the form of competitive ranks or hierarchy. In parallel to this, we also developed a strict definition for aggression, under which we only included traits if they described the intensity of antagonistic behaviors and not the formed social hierarchy. Based on the narrow- and broad-sense definitions, therefore, we built 2 complementary data sets. When relying on more permissive definitions, we could test our predictions based on a larger sample size in terms of the number of studies for each pair-wise correlation. On the contrary, the stricter criteria allowed us to achieve higher control for the variation in measurement conditions on the cost of lower sample size. Hence, we performed our analysis with both data sets.

From the collected papers, we extracted data on the magnitude and direction of the relationship between any 2 of the above-defined variables in a form of correlation coefficient as a standard measure of effect size. We used the program Comprehensive Meta-analysis (Borenstein 2010) to calculate effect sizes based on the conventional formulas (Cohen 1992; Walker 2003; Nakagawa and Cuthill 2007). If results were provided in a nonstandard way, for which the program does not provide transformation method, we used procedures described in Garamszegi, Markó, et al. (2012) to derive effect sizes. We determined the direction of the relationships according to the expectations from the intensity levels of behaviors that predict positive relationships among behavioral traits. This does not always mean that the measured activity, exploration, risk-taking, and aggression variables show positive correlation, because approach distances and latencies are inverse measures of intensity. In such cases, the sign of the relationship was converted appropriately. If it was necessary, we corresponded with the original authors to clarify uncertain issues in association with their analyses that we needed to use for calculating effect sizes.

We identified data as independent based not on whether they were published in the same scientific paper, but by a lack of overlap of studied individuals. We dealt with the nonindependence of data by either careful data sorting or by using hierarchical models during statistical analyses. At the level of data base building, if a source paper provided results separately for different sets of individuals (e.g., for different species, sexes, or age categories), we handled them as independent studies and included the associated effect sizes as if they had originated from different studies. In contrast, if different papers by the same authors relied on the same (or on an overlapping sample of) individuals, we combined the derived effect sizes into a single, study-wise effect size. In some cases, the investigators measured more than 1 variable for the same behavioral trait (e.g., approach distance, latency movement intensity) within the same study. Therefore, in these instances, we could obtain multiple correlations that actually reflected the same relationship in essence. To avoid pseudoreplication in these cases, we lumped these homologues within the study level by calculating the weighted average of the calculated effect sizes for the same relationship. The application of our definitions and data sorting strategy resulted in 191 (broad-sense definitions) and 104 (strict definitions) entries in the 2 data sets. At the level of analyses, we dealt with the following cases of nonindependence. In a considerable number of papers, more than 1 effect size could be used, as these sources provided multiple correlations for different behavioral traits (e.g., between aggression and risk-taking and between novel object exploration and risk-taking), which cannot be treated as being independent as they relied on the same individuals. In these cases, we created an "outcome" categorical variable to reflect the type of comparison based on the 15 possible pair-wise combinations of activity/aggression/novel environment exploration/novel object exploration/risk-taking and included it in our statistical model to describe the hierarchical organization of data. As different study outcomes from the same paper correspond to higher similarity in the methodology, whenever we used multiple entries from the same study (i.e., correlations for different pair-wise comparisons), we also treated such nonindependence via hierarchical modeling of study-specific effects (see more details below).

We also collected data on the repeatability of traits if these were presented in the source paper (Supplementary Table S2). Repeatability can be calculated by various methods, such as by an ANOVA-based estimation of the proportion of the between-individual variance relative to the total variance (Lessells and Boag 1987) or by an intraclass correlation coefficient (Sokal and Rohlf 1995), which are inconsistently applied across studies due to different constraints of the underlying sampling. For simplicity, we combined different estimates of repeatability irrespective of the methods used, by adopting the convention that negative values imply 0 repeatability, thus forcing the entered estimate to vary between 0 and 1.

Hierarchical meta-analysis models: model construction concepts

We performed our analyses in R statistical environment (R Development Core Team 2007) with different packages. For the analyses, we transformed r to its normalized form Fisher's $\mathcal{Z}$ (Fisher 1915). Because we assumed considerable variability in the effect sizes across studies as they corresponded to different species or conditions, we consistently relied on random-effects models for all tests.

We considered the potentially confounding role of nonindependence of data as caused by 1) the use of multiple effect sizes from the same studies, 2) the use of multiple effect sizes from the same species, and 3) the phylogenetic relatedness of species (Nakagawa and Santos 2012). To overcome these problems, we applied hierarchical meta-analysis models, which can effectively partition correlation structures within different levels (e.g., within studies and species) and can also take into account the phylogenetic relatedness of species. For such a statistical exercise, we followed the Bayesian quantitative genetic approach that is available in the MCMCglmm package (Hadfield and Nakagawa 2010) and that permits phylogenetic mixed models based on Markov chain Monte Carlo (MCMC) algorithms. In this framework, we first created different sets of models with different random-effect structures (e.g., study source or species) and with or without addition of phylogenetic variance components and different predictors (see model specifications in Tables 1 and 3). For the phylogenetic models, we used the ultrametric phylogenetic tree of species from Garamszegi, Markó, et al. (2012) with branch lengths being scaled to unit length (distance from root to tip). We investigated if effect sizes among pair-wise comparisons are different by entering the outcome category as a fixed effect. The influence of moderators was also assessed by using the appropriate terms in the model description (e.g., spatial or temporal overlap). Unfortunately, data were limited to run models with combined random effects (e.g., study + species, as for many species there was only one study, leading to highly unbalanced design); thus, we were constrained to use models with a single random term for the assessment of data fit. However, we could apply model-averaging techniques to simultaneously evaluate the combined effects of the random factors on the parameter estimates (i.e., effect sizes and CIs) based on the relative fit of the models including such terms (Symonds and Moussalli 2011). As a measure of model fit, we relied on the deviance information criterion (DIC), with the lowest value offering a better fit to the data. Models were ranked along their DIC estimates and we computed $\exp(-\mathrm{DIC}/2)$ for subsequent model averaging (Congdon 2005). We derived the estimates of effect sizes and their CIs for different groups of comparisons following Schielzeth (2010) and weighted them across models based on relative DIC analogously to procedures relying on AIC weights (Symonds and Moussalli 2011).

Given that we had no preceding information, we used noninformative priors throughout (V = 1E-10, nu = −1). The Markov chains were allowed to run up to 13 000 iterations, with 3000 iterations of burn-in and with 10 iterations of thinning interval. Convergence diagnostics based on Gelman and Rubin statistics (Gelman and Rubin 1992) performed over chains from repeated runs of the same model unanimously indicated that this sampling scheme was appropriate to achieve convergence. By examining outputs from the repeated runs, we also checked the stability of results and assessed the influence of alternative prior considerations.

Our personal experience based on these repeats is that although MCMCglmm offers flexible statistical designs for meta-analyses that take into account phylogenetic associations and other hierarchical structures present in the data based on Bayesian approaches, its practical applicability for complex models is undermined by the instability of the results. We detected that different runs, even after proper convergence, can give somewhat different results in terms of both model fit and parameter estimates, if too many factors were considered in the model. In addition, results appeared to be sensitive to different prior settings with unknown biological meaning. Therefore, we present results that are averaged across runs and that correspond to the noninformative priors. Moreover, we avoided obtaining parameter estimates from very complex models (e.g., when moderators were also included) in MCMCglmm, and we also verified our results by using an alternative approach based on the R program metafor (Viechtbauer 2010). The metafor package is a collection of general functions that allow performing fixed- and random-effects meta-analyses to obtain main effect and heterogeneity, carrying out tests of moderator variables and meta-regressions, and assessing publication bias. The shortcoming of this program is that it cannot deal with phylogenetic effects and with the hierarchical organization of data. Therefore, we used metafor and MCMCglmm complementarily to balance between their benefits and weaknesses.

Pair-wise relationship between traits

First, we focused on the potential differences in mean effect sizes among outcomes. Hence, from the above models, following the model averaging of parameters, we determined the relationship for each possible pair-wise combination of the 5 behavioral traits (e.g., activity and novel environment exploration, activity and novel object exploration, activity and risk taking, activity and aggression, novel environment exploration and novel object exploration, etc.). For the data set that corresponded to the broader definitions, we could also derive mean effect size statistics for symmetric relationships. This was because the broader sense definitions allowed us to include associations between the same behavioral traits that were defined under different conditions (e.g., because in the same study, aggression could be defined based on both aggressive interactions and competitive ranks, or risk-taking could be defined based on behavioral response to both predator and other threat stimuli). Therefore, we present correlations obtained from our more broadly encompassing variable definitions that corresponded to a larger sample size but with lesser precision (Tables 1 and 2). For comparison, we also provide the main figure derived from the data set with our stricter definitions, with a smaller sample size but higher precision (Supplementary Tables S3 and S4). We report all effect sizes in the form of correlation coefficients (r) together with the associated 95% CIs. In ecology, the following benchmark is used for interpretations: $r \approx 0.1$ is a small effect, $r \approx 0.3$ is a medium effect, and $r \approx 0.5$ is a strong effect (Cohen 1988; Møller and Jennions 2002).

Table 1

MCMC modeling of different grouping, random-effect structures, and phylogenetic effects that were built in the meta-analysis models when using broad-sense criteria for variable definition

Model	Random terms	Grouping variable	Phylogeny	DIC
1)	—	—	—	100.95
2)	Study	—	—	15.35
3)	Species	—	—	52.34
4)	Species	—	+	92.51
5)	—	Outcome	—	112.68
6)	Study	Outcome	—	−66.57
7)	Species	Outcome	—	35.57
8)	Species	Outcome	+	91.69

Grouping variable was used in the MCMCglmm exercise as a moderator variable.

Based on the estimated pair-wise correlations, we also computed the partial correlations between traits by using the corpcor library in R (Schaefer et al. 2012). Partial correlation measures the degree of association between 2 variables when the effect of a set of other variables is held constant (Sokal and Rohlf 1995). If different traits collectively formed a syndrome through correlations of similar magnitude, we expected to find comparable partial correlations as well. However, if certain traits appear associated with others through another variable (which has no relevance for behavioral syndromes), we predicted that partial correlations would be different. For example, if a detected relationship between aggression and risk-taking is principally mediated by novel object exploration (i.e., how individuals respond to novelty situation of any kind), this would be observed by their weak partial correlation that is independent of novel object exploration.

For each relationship, we also aimed to perform tests of heterogeneity (DerSimonian and Laird 1986) and of publication bias based on funnel plot analysis and the imputation of missing studies (Begg and Mazumdar 1994; Duval and Tweedie 2000), which are unavailable in MCMCglmm. We used metafor for such purposes. In these exercises, we selected the appropriate subset of the data based on the outcome variable to focus on a single comparison only from each study, ensuring that the sample used to calculate the mean statistics involved independent entries. We repeated this procedure for each possible pair-wise correlation of the 5 behavioral traits. Heterogeneities are presented in Table 2, whereas estimates of publication bias are reported in Supplementary Table S5 and Figure S1, which generally show weak evidence for biased publishing.

Effect of moderators: temporal and spatial overlaps

We tested for the effect of contextual overlap on the mean effect sizes by using 2 moderator variables. Spatial overlap was a bivariate grouping variable, with “same” indicating whether the 2 assays were performed in the same experimental compartment and “different” indicating whether the 2 behaviors were characterized in spatially separated setups (e.g., in the same or in different test chambers,

Table 2

The general relationships among behavioral traits when relying on broad-sense definitions (see Methods)

	Activity	Aggression	Novel environment	Novel object	Risk-taking
Activity	$r = 0.492$ $(-0.040/0.764)$, $P = 0.054$, $N_{\text{studies}} = 1$, $Q = 0$	$r_{\text{partial}} = 0.062$	$r_{\text{partial}} = 0.281$	$r_{\text{partial}} = -0.099$	$r_{\text{partial}} = 0.185$
Aggression	$r = 0.133$ $(-0.004/0.286)$, $P = 0.088$, $N_{\text{studies}} = 18$, $Q = 53.879$***, $I^2 = 74.34\%$	$r = 0.432$ $(0.210/0.610)$, $P < 0.001$, $N_{\text{studies}} = 6$, $Q = 68.642$***, $I^2 = 92.47\%$	$r_{\text{partial}} = 0.089$	$r_{\text{partial}} = 0.198$	$r_{\text{partial}} = 0.067$
Novel environment	$r = 0.345$ $(0.163/0.497)$, $P = 0.002$, $N_{\text{studies}} = 9$, $Q = 33.934$***, $I^2 = 74.97\%$	$r = 0.203$ $(0.074/0.323)$, $P = 0.004$, $N_{\text{studies}} = 23$, $Q = 49.495$***, $I^2 = 57.78\%$	$r = 0.348$ $(0.127/0.551)$, $P = 0.008$, $N_{\text{studies}} = 5$, $Q = 21.564$***, $I^2 = 87.95\%$	$r_{\text{partial}} = 0.161$	$r_{\text{partial}} = 0.190$
Novel object	$r = 0.074$ $(-0.114/0.268)$, $P = 0.466$, $N_{\text{studies}} = 10$, $Q = 7.201$, $I^2 = 0\%$	$r = 0.273$ $(0.093/0.419)$, $P = 0.002$, $N_{\text{studies}} = 14$, $Q = 43.374$***, $I^2 = 75.59\%$	$r = 0.280$ $(0.134/0.402)$, $P < 0.001$, $N_{\text{studies}} = 17$, $Q = 27.673$*, $I^2 = 42.13\%$	$r = 0.265$ $(-0.013/0.522)$, $P = 0.088$, $N_{\text{studies}} = 7$, $Q = 15.870$*, $I^2 = 63.38\%$	$r_{\text{partial}} = 0.322$
Risk taking	$r = 0.270$ $(0.094/0.432)$, $P < 0.001$, $N_{\text{studies}} = 12$, $Q = 48.021$***, $I^2 = 72.66\%$	$r = 0.208$ $(0.104/0.319)$, $P < 0.001$, $N_{\text{studies}} = 28$, $Q = 47.899$**, $I^2 = 44.52\%$	$r = 0.347$ $(0.239/0.448)$, $P < 0.001$, $N_{\text{studies}} = 28$, $Q = 99.470$***, $I^2 = 75.98\%$	$r = 0.397$ $(0.244/0.533)$, $P < 0.001$, $N_{\text{studies}} = 11$, $Q = 15.937$, $I^2 = 39.36\%$	$r = 0.566$ $(0.250/0.783)$, $P = 0.004$, $N_{\text{studies}} = 2$, $Q = 0.241$, $I^2 = 0\%$

Below the diagonal, mean effect sizes are presented together with the associated 95% CIs (lower/upper) and sample sizes (number of studies) as obtained from the Bayesian modeling of different levels of nonindependence. Parameters are model-averaged estimates from models 4–8 in Table 1. Q statistics are for the tests of heterogeneity as obtained from metafor when applied separately to different trait combinations. Above the diagonal, partial correlations that are calculated based on the mean effect sizes are shown. The diagonal represents associations among the same behavioral traits that were defined under different conditions.

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

tanks, rooms, or territories, respectively). Temporal overlap was a 5-state variable, with “0” standing for tests that were made immediately after each other, “1” for tests made with some pause interval between them but on the same day, “2” for tests made on different days but in the same season/life cycle, “3” for situations in which tests were made in different seasons/life cycles but in the same year, and “4” for tests that were made in different years. As intermediate states are meaningful, we treated this variable as a continuous predictor. Given that the quantitative differences between states are different (spanning from minutes to years), the attempt to treat them on a continuous axis is meaningful on a logarithmic scale.

To examine if these 2 estimates of the contextual overlap of the assay conditions can affect effect sizes of the particular relationships, we added these moderators as predictor variables to the hierarchical models with different random effect structures (Table 4). After fitting these models using MCMCglmm, we chose the one that resulted in the lowest DIC, and for the stability issues detailed above, we created an analog model in metafor to obtain group-specific parameter estimates. This approach also permitted us to assess the significance of the between-group comparisons (e.g., for the difference between “same” and “different” overlap categories) within each pair-wise combination of behaviors.

Repeatability

Repeatability was also analyzed in a meta-analytic framework, where it was treated as a correlation (r) to estimate effect size. However, given that Fisher’s Z transformation satisfies criteria for variance stabilization and normality only when the number of repeats is $k = 2$, we applied the Konishi–Gupta modified Z transformation that is more effective in cases of $k > 2$ (Konishi and Gupta 1987). We considered hierarchical structuring at the within- and between-study level; thus, we used the appropriate random effect term in the statistical modeling. As we were interested to see whether repeatability (measurement error in a statistical sense, but in a biological sense, it can involve important intraindividual variations) varied among activity, aggression, novel environment exploration, novel object exploration, and risk-taking, we tested if adding “trait” as a fixed factor improved model fit in the MCMCglmm exercise. We also examined the effect of this moderator in metafor to check if variation of effect size among traits was a significant source of the overall heterogeneity in the repeatability data.

RESULTS

The constructed models and their data fit statistics when using broader definitions and ignoring contextual overlap are given in Table 1 (see Supplementary Table S3 for analogous results when using strict definitions). DIC estimates suggest that the most important confounding effects that need to be considered are the nonindependence of data that appear within the same study and within the same species. However, although species-specific effects seem to be important, the Bayesian exercise clearly indicated that these are independent of phylogenetic history of species, because models considering phylogenetic relatedness failed to improve the fit to the data. When the nonindependence of data within the same study and species is taken into account, differentiating effect sizes among different outcomes results in further increase in model fit, suggesting that effect sizes are heterogeneous across different comparisons of behavioral traits. Model-averaged effect sizes for each pair-wise relationship based on the relative fit of models are given in Table 2 and Supplementary Table S4. Note that the first-best model was the one that controlled for study-specific effects and had a DIC value that was considerably lower than that of the second-best model; thus, model-averaged parameters basically represent the outcome of the first-best model.

Pair-wise relationships between traits

In general, we found a modest but consistent positive relationship between almost all pair-wise combinations of behavioral traits (Table 2, Supplementary Table S4, and Figure 1). The mean effect size varied between 0.074 and 0.566 when the variables were defined based on our broader criteria (Table 2) and between 0.071 and 0.442 when we relied on stricter definitions for the variables (Supplementary Table S4). When a relatively large number of studies were available for a particular comparison, the 2 approaches gave similar results. The only considerable difference that emerged was for the activity/risk-taking relationship, as variables defined based on our broad-sense criteria revealed a mean effect size of 0.270, whereas the use of strict definitions led to a nonsignificant mean effect size of 0.071. However, only 2 studies could be used for the latter comparison, which makes conclusions for this relationship uncertain. Given the similarity of the 2 sets of results when reasonable sample sizes were available, for further interpretations and analyses, we relied on the results that correspond to the broader definitions and to a larger sample size.

Few studies presented associations among the same behavioral traits that were defined under different conditions by different variables (diagonal in Table 2). Considering the wider CI around these estimates, the symmetric relationships did not seem to provide consistently stronger relationships than the asymmetric relationships, that is, the pair-wise correlations of different behavioral traits (below the diagonal in Table 2 and Figure 1). In other words, the relationship among different behavioral variables describing the same phenomenon is similar to the relationship among behavioral variables corresponding to different ecological situations.

Inspecting the particular effects (Table 2 and Figure 1), the correlation between activity and novel environment exploration appeared stronger than the correlation between activity and novel object exploration (more than 4-fold difference with a relatively little overlap among 95% CIs). In comparison, risk-taking and aggression were related to both measures of exploration with a similar magnitude (Table 2 and Supplementary Table S4). In these latter cases, we infer that the difference in the significance of these estimates is a question of sample size and the precision of the estimate and therefore cannot be used to make strong inferences about the magnitude of the effects.

Based on the mean effect sizes estimated for each pair-wise relationship, we also calculated the partial correlations among traits in order to hold constant the covariation among variables. This control revealed that aggression is more likely associated with novel object exploration than with novel environment exploration (2-fold difference) when the association between the 2 estimates of exploration (r = 0.280) is factored out. Such an asymmetry for the 2 exploration traits is also prevalent in relation to activity, as novel environment exploration shows a positive and stronger relationship with activity than novel object exploration, which in fact reveals a negative tendency. Notably, aggression shows a generally weak partial relationship with other traits (<0.1) with the exception of novel object exploration. Consequently, both the mean effect sizes calculated for each comparison and their partial correlations indicate that the correlation structure is not symmetric across behavioral traits.

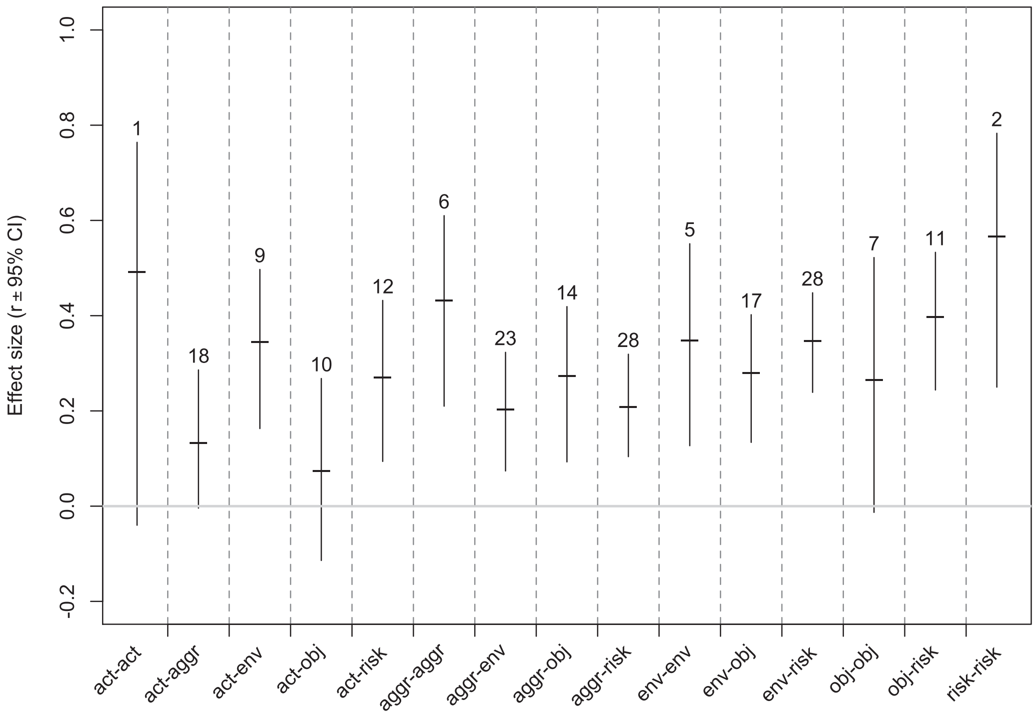

Figure 1
Effect sizes for each pair-wise relationship among behavioral traits when relying on broad-sense definitions (see Methods). Means (horizontal thick marks) and 95% CIs (error bars) are calculated from multilevel meta-analyses presented in Tables 1 and 2. Numbers indicate sample sizes in terms of the number of studies. act: activity, aggr: aggression, env: novel environment exploration, obj: novel object exploration, and risk: risk-taking.

Effect of moderators: temporal and spatial overlap

Based on a hierarchical modeling of different random effects in combination with the interaction between outcome and contextual overlap in the Bayesian framework, we found that the most important factor that can cause nonindependence problems is the common origin of data from the same study, and the consideration of phylogenetic effects did not improve model fit (Table 3). Therefore, we avoided obtaining parameter estimates from unnecessarily complex models that simultaneously consider moderators and phylogeny but proceeded with the simpler, metafor-based approach under the assumption that phylogeny does not inflate the results.

In that framework, we estimated the relationship between contextual overlap and effect sizes with a single entry from each study and repeated these procedures for each pair-wise relationship separately (i.e., to remove nonindependence caused by study-specific effects). Concerning the spatial overlap between the experimental conditions, we found that the relationships between aggression and novel object exploration, between activity and risk-taking, and between novel environment exploration and novel object exploration were systematically stronger when the 2 behaviors under comparison were scored in the same experimental compartment (Table 4, below the diagonal, and Figure 2). These dissimilarities involved 2- to 5-fold differences between the "same" and "different" groups of effect sizes. However, when we investigated the influence of temporal overlap between assay conditions, we found no significant evidence for this moderator to be an important determinant of the detected relationships among behavioral traits (Table 4, above the diagonal). Therefore, our predictions about the importance of contextual overlap mediating the detected relationship among behaviors received partial support.

Table 3

MCMC modeling of spatial and temporal overlaps when different random effects were also considered

Model	Random terms	Grouping variable	Phylogeny	DIC
Model set for spatial overlap				
1)	—	—	—	100.24
5)	—	Outcome	—	112.68
9)	—	Spatial overlap	—	99.24
10)	Study	Spatial overlap	—	21.68
11)	Species	Spatial overlap	—	50.03
12)	Species	Spatial overlap	+	94.22
13)	—	Spatial overlap × outcome	—	92.74
14)	Study	Spatial overlap × outcome	—	−53.83
15)	Species	Spatial overlap × outcome	—	38.33
16)	Species	Spatial overlap × outcome	+	86.67
Model set for temporal overlap				
17)	—	—		90.01
18)	—	Outcome	—	104.22
19)	—	Temporal overlap	—	92.41
20)	Study	Temporal overlap	—	1.34
21)	Species	Temporal overlap	—	39.97
22)	Species	Temporal overlap	+	80.45
23)	—	Temporal overlap × outcome	—	104.71
24)	Study	Temporal overlap × outcome	—	−94.24
25)	Species	Temporal overlap × outcome	—	17.95
26)	Species	Temporal overlap × outcome	+	65.42

Grouping variables were used in the MCMCglmm exercise as moderator variable (discrete or continuous predictor). Models with × indicate models in which the predictors were added in a full combination (i.e., main effects and interaction terms). DIC values within the model set for spatial overlap (models 9–16) can be compared with models in Table 1 (models 1–8), as they rely on the same data. However, due to some missing data for temporal overlap, models in the lower set (models 17–26) cannot be compared with other sets as model fit statistics do not correspond to the same data.

Repeatability

In the Bayesian framework, the null model disregarding any hierarchical structure offered a worse fit to the data than the model that considered variation among and within studies (DICs: 238.47 vs. 170.02). However, adding "trait" as an additional predictor to the model did not result in further improvement (DIC: 210.76). Similarly, the effect of this moderator was statistically nonsignificant in the corresponding metafor model ($Q = 8.704$, df = 4, $P = 0.069$). Trait-wise mean estimates show that repeatability varies among traits within a range of 0.29 and 0.46, with a substantial overlap in the associated CIs (Table 5). Note that these repeatability estimates differ from the correlations presented in the diagonal of Table 2. The repeatabilities in Table 5 are calculated based on exactly the same traits measured at different times, whereas the correlations in Table 2 are for different behavioral traits that essentially represent the same phenomenon.

DISCUSSION

The most important findings of this meta-analysis are that 1) the general phenotypic relationships among behavioral traits were moderate and varied across different pair-wise comparisons, 2) the correlations between the same behavioral traits assayed in different ways were similar to the correlations among different behaviors, 3) novel object exploration and novel environment exploration correlated only moderately and showed asymmetric correlations with activity, 4) spatial overlap between the conditions of the behavioral assays had an effect on the detected correlation, as certain effect sizes were higher when the behavioral traits under comparison were estimated in the same experimental compartment, and 5) repeatability of the studied behavioral traits was statistically indistinguishable. For the interpretations of our results, in the light of recent discussions (Dingemanse et al. 2012; Garamszegi and Herczeg 2012; Garamszegi, Markó, et al. 2012), we assume that phenotypic correlations reflect correlations among individual-specific behaviors and that the within-individual correlations are negligible.

The magnitude of effects

The possible reasons of finding moderate effects for behavioral correlations in general are discussed elsewhere (Garamszegi, Markó, et al. 2012), and such arguments also hold here. Briefly, the fact that the phenotypic relationship among behaviors is not particularly strong could mean that 1) behaviors are subject to considerable within-individual variations or/and can only be measured with considerable errors, leading to low repeatability; 2) the strong genetic relationship among traits is masked by moderate heritability; 3) different behaviors weakly correlate because a substantial number of individuals fail to conform to the

Table 4

The effect of contextual similarity (spatial and temporal overlaps) between the conditions of the 2 behavioral assays on the relationship between the 2 measured traits

	Activity	Aggression	Novel environment	Novel object	Risk taking
Activity		β (±SE) = −0.070 (0.092), df = 1,15, Q = 0.578, P = 0.447	β (±SE) = −0.165 (0.380), df = 1,6, Q = 0.189, P = 0.664	β (±SE) = 0.045 (0.077), df = 1,7, Q = 0.338, P = 0.561	β (±SE) = −0.091 (0.161), df = 1,9, Q = 0.320, P = 0.571
Aggression	df = 1, N_{studies} = 18, Q = 0.451, P = 0.502		β (±SE) = 0.161 (0.116), df = 1,20, Q = 1.179, P = 0.167	β (±SE) = −0.141 (0.207), df = 1,11, Q = 0.461, P = 0.497	β (±SE) = −0.022 (0.055), df = 1,25, Q = 0.164, P = 0.686
Novel environment	df = 1, N_{studies} = 9, Q = 0.080, P = 0.777	df = 1, N_{studies} = 23, Q = 0.727, P = 0.394		β (±SE) = 0.029 (0.121), df = 1,14, Q = 0.057, P = 0.811	β (±SE) = 0.061 (0.061), df = 1,25, Q = 0.981, P = 0.322
Novel object	df = 1, N_{studies} = 10, Q = 1.057, P = 0.310	df = 1, N_{studies} = 14, Q = 3.889, P = 0.049	df = 1, N_{studies} = 17, Q = 8.449, P = 0.004		β (±SE) = −0.041 (0.075), df = 1,9, Q = 0.293, P = 0.588
Risk-taking	df = 1, N_{studies} = 12, Q = 5.933, P = 0.015	df = 1, N_{studies} = 28, Q = 0.027, P = 0.868	df = 1, N_{studies} = 28, Q = 1.223, P = 0.269	df = 1, N_{studies} = 11, Q = 0.001, P = 0.998	

Tests for moderator-mediated effects using metafor when spatial overlap was a grouping predictor (lower diagonal) and when temporal overlap was a continuous predictor and was used in meta-regression (upper diagonal). For better interpretation, parameter estimates (β) and their standard errors (SE) are only given for the continuous moderator (temporal overlap), and effects for the categorical predictor (spatial overlap) are presented in Figure 2.

patterns predicted by the syndrome; and 4) there is considerable variance in effect sizes among studies (i.e., a few particular studies find strong effects, whereas others mostly report weak effects) due to differences in methodology or in the ecological background. Some evidence already exists to support these possibilities (e.g., modest repeatability: Bell et al. 2009; modest heritability: van Oers et al. 2004; differences in methodology: Burghardt et al. 2012; Garamszegi, Markó, et al. 2012; and differences in ecology: Dingemanse et al. 2007). This study further suggests a role for the effect of study design (Table 4), and the fact that different variables can be measured by moderate repeatability (Table 5) points to the importance of measurement error or intraindividual variation. In any case, the general consequence of these different mechanisms mediating the observed patterns of pair-wise phenotypic correlations cannot be teased apart based on the correlative nature of available data. The only conclusion we can make is that the generally weak relationship among behaviors we found in Garamszegi, Markó, et al. (2012) is unlikely to be caused by the variance in the degree of relatedness among particular behaviors, as each comparison appeared to be moderate at the best.

From the practical point of view, the widely detected weak phenotypic correlation among traits suggests a high probability of committing type II errors (i.e., to fail to reject the null hypothesis of no syndrome structure when it is false) when correlations are based on modest sample sizes. Although we did not find evidence for publication bias in the current study (Supplementary Table S1 and Figure S1), the possibility of committing type I errors (i.e., to reject the null hypothesis of no syndrome structure when it is true) cannot be safely excluded either. Our previous appraisal based on a combined data set (thus relying on a larger sample size) would seem to suggest that a considerable number of studies reporting smaller effect sizes may remain unpublished (Garamszegi, Markó, et al. 2012). Type I errors are likely to be present in the "personality" and "syndrome" literature, because studies are usually based on moderate sample sizes that can lead to overestimates of effect sizes, especially when the true effect size is small (Ioannidis 2005; Gelman and Weakliem 2009). Based on these statistical considerations and on the fact that the relationships among behaviors are likely to be modest or small, we suggest that future studies of behavioral syndromes would benefit from focusing on estimates of effect sizes and CIs instead of significance levels (sensu Nakagawa and Cuthill 2007). Such an approach will bring continuous measures of trait association to the forefront of our research and will help abandon the misleading binary thinking about behavioral syndromes.

Variance of effect sizes across comparisons

The most important finding of this study (focusing separately on each pair-wise comparison) is that, in addition to heterogeneities within particular comparisons, there are remarkable differences in effect sizes across comparisons as well. The best data fit offered by the model that discriminates among outcomes, while simultaneously controlling for nonindependence within studies (Table 1), clearly indicates that pair-wise correlations cannot be treated as estimates of the same general correlation that exists among all behaviors. The most obvious discrepant results appeared in association with the 2 exploration traits, which are supposed to reflect the same phenomenon (Réale et al. 2007), but which showed only moderate correlation with each other and a different degree of association with activity. Furthermore, we also found asymmetric partial correlations among behaviors (ranging from −0.099 to 0.322) after controlling for the covariance among them. Noticeably, aggression

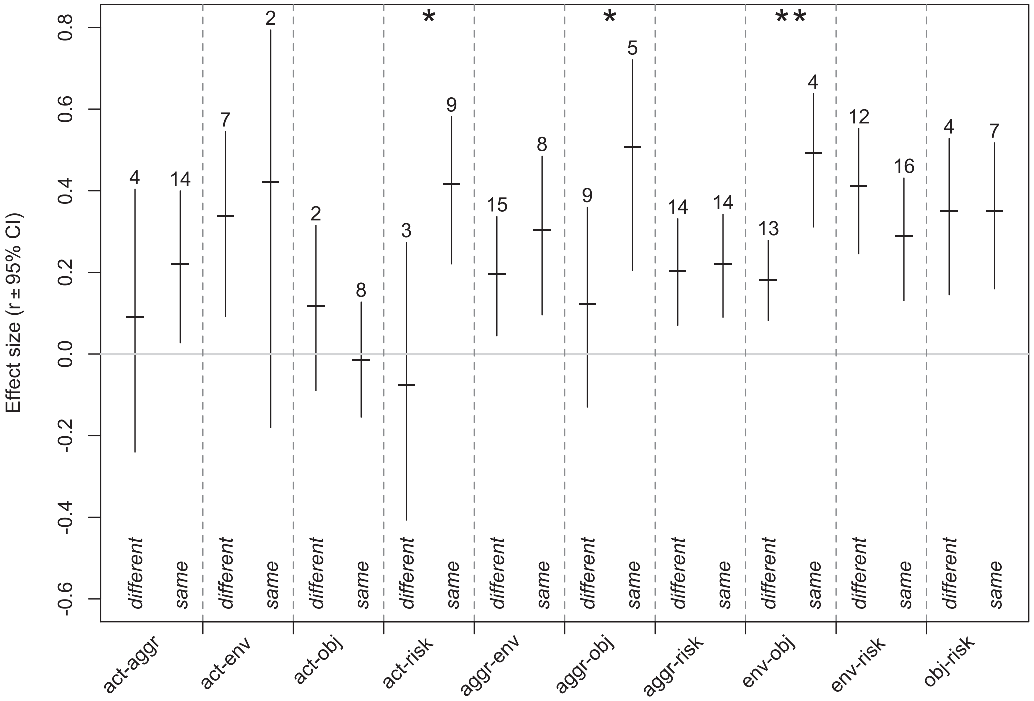

Figure 2

Means (horizontal thick marks) and 95% CIs (error bars) of effect sizes (*r*) for the relationship among particular behavioral traits when tabulated separately for experimental setups with and without spatial overlap between the 2 behavioral assays (different: the 2 traits were estimated in different experimental compartments; same: the 2 traits were estimated in the same experimental compartment). Numbers indicate sample sizes in terms of the number of independent studies. act: activity, aggr: aggression, env: novel environment exploration, obj: novel object exploration, and risk: risk-taking. **P* < 0.05, ***P* < 0.01 as given in the corresponding test (see Table 4).

Table 5

Repeatability estimates of traits and the associated 95% CIs based on MCMC modeling when using study as random factor (values are back-transformed to the original scale)

Trait	Repeatability	95% CI_{lower}	95% CI_{upper}
Activity	0.457	0.231	0.641
Aggression	0.451	0.212	0.665
Novel environment	0.289	0.019	0.572
Novel object	0.435	0.200	0.687
Risk taking	0.362	0.139	0.602

showed a considerable level of association with novel object exploration (~0.2), whereas it was only weakly related to other behavioral traits including novel environment exploration.

The variation in the relationship between different pairs of behaviors may indicate that certain associations are under different biological controls. For example, selection can favor stronger or weaker correlations among particular pairs of behaviors in different circumstances, resulting in population differences in the optimal trait combinations (e.g., Dingemanse et al. 2007). The presence of a predator or the limitation of resources may shift the adaptive value of risk-taking, aggression, or exploration, which can have consequences for the trade-off among these traits, resulting in greater variance in the particular relationships in different environments (Sih, Bell, & Johnson 2004; Sih, Bell, Johnson, & Ziemba 2004). The generally weak relationship among traits may partly result from the fact that there is a substantial variation among studies in terms of findings, and the pooled data combine all types of effects in which strong positive and negative effects extinguish each other. If such an extensive variation among studies exists, this may point to the importance of the underlying ecological background and the possibility that different associations are under different biological controls.

The differences in effect sizes between outcomes and studies may indicate a weak role for the common government of behaviors. One can hypothesize that behavioral correlations emerge because the particular traits are mediated by the same innate characteristic of individuals. For example, temperament can be such a background trait, which would imply that individuals with shy attitude display less active behaviors in general than bold individuals (Buss et al. 1987; Clark and Wilson 1999; Gosling 2001). If the manifestation of such a characteristic occurs similarly in each behavior, we can predict that the relationship among different traits will be similar, which was not the case in this study. However, we infer that the rejection of the temperament concept would be premature at the current state of research, because we could not explain all components of heterogeneity. Although we considered issues about contextual overlap, repeatability (see below) and species-specific effects, based on the correlative nature of the study, we could not control for every confounder that can potentially raise differences in the detected correlations even if they are mediated by the same background trait. Future studies testing for the effect of other mediators may be more successful in fully explaining differences in effect sizes. If the difference between the mean behavioral correlations still exists after considering all confounders, one can safely reject the temperament concept.

Note that measurement errors and repeatabilities may also render correlations different, if these differ among assays of different behaviors, resulting in traits with lower repeatability generally showing weaker relationships. However, relying on the available sample, we were unable to prove that repeatability varies across traits. There was a weak tendency for exploration of a novel environment to be estimated with lower repeatability than other traits. If this were the case, we should have detected generally weaker relationships for this trait when correlating with others, but this was not the case (Tables 2 and 4). The fact that usually the same traits (e.g., movement activity, approach distance, or latency) are measured in the behavioral assays of different traits may explain why the corresponding repeatabilities cannot be discriminated. As we found no sign of particular behaviors being more flexible than others, similarity in repeatability may indicate that individuals are generally similar in their consistency among the studied domains.

Although this study does not offer detailed insights on the underlying evolutionary or physiological mechanisms, the asymmetries in the correlations we detected imply that the studied main domains are not necessarily and equally independent of each other. Therefore, treating all behavioral traits similarly for the syndrome concept might be deceptive. In a particular study's finding, for instance, a relationship between activity and novel environment exploration may have a different meaning than a relationship between activity and novel object exploration. The former association may be more likely to appear because activity and novel environment exploration reflect similar phenomena (i.e., general activity in any environment), whereas the latter correlation may be more likely to signify that different domains are under the same selection pressure (activity and fearfulness to novelty). The 2 types of results have different implications for the concept of behavioral syndromes. Novel object exploration and novel environment exploration are often combined as 2 components of exploration (e.g., Verbeek et al. 1996; Wilson and Godin 2009; Mafli et al. 2011). This practice may be misleading if these 2 variables have different biological meanings.

Contextual overlap

Even if behaviors respond to selection similarly, they may still be confounded by contextual overlap, which can vary from one pair-wise correlation to another, raising dissimilarities in the detected relationship. We found some specific evidence for this in that certain behavioral correlations might only be detectable because of the overlap of the corresponding contextual situation. Namely, the relationships between novel object exploration and aggression, between novel object exploration and novel environment exploration, and between activity and risk-taking are systematically weaker when the underlying behavioral variables were obtained in different experimental compartments (Figure 2). This result might reflect at least 2 different mechanisms. First, it is possible that some correlations emerge due to individual variations in the stress/novelty-coping ability to the experimental environment, which affects all measured behaviors in a similar way, and not necessarily because the behaviors form a syndrome that resulted from a long-lasting selection process (see Maier et al. 1988; Budaev 1997; Réale et al. 2007; Garamszegi et al. 2009). Second, it might simply mean that by measuring different traits in different environments, researchers contradict the basic principle of common garden/standardized experimental studies by introducing uncontrolled environmental variation that hampers the actual pattern of interest. Therefore, we suggest that (unless one measures novel environment exploration) proper acclimation times should be allowed for the test animals in the test environments before the test stimulus is presented, and if different behaviors are measured in different environments for various reasons, behaviors of a subsample of individuals should be measured in each environment to separate consistency (i.e., individual-specific value) from the effect of the environment.

The confounding role of contextual overlap raises issues for the proper categorization of behavioral traits. Réale et al. (2007) suggested domains to be discriminated according to environmental situation in which the behavior is displayed. This categorization inherently assumes that focal traits are independent units and can be separated from each other. Although such independence may be warranted theoretically, this assumption is not necessarily met in practice, because as we showed here, certain degrees of contextual overlap may exist among behaviors. Therefore, when relying on the definitions of Réale et al. (2007), researchers may also wish to deal with the confounding role of the overlap between the contextual backgrounds of focal traits. This is crucial for the interpretation of the detected behavioral correlation, because such a relationship can indicate the existence of both "true" behavioral syndrome and contextual overlap.

Phylogenetic signal and heterogeneity

Two coherent patterns emerged in most of the analyses: the weak role of phylogenetic effects and the relatively high heterogeneity that accompanied the mean effect size estimates. The lack of phylogenetic signal in the data may indicate that effect sizes are not conserved evolutionarily and thus are essentially independent (Chamberlain et al. 2012). Flexible behavioral traits (and their correlations) may be subject to relatively large variations that likely hinder the phylogenetic constraints that act on them (Blomberg et al. 2003) and that result in even closely related species demonstrating remarkably different syndrome structures. It is also plausible that we failed to detect phylogenetic effects because other factors that vary independently of phylogeny (e.g., methodological differences) have such a strong impact that they override the impact of common ancestry. However, the simultaneous consideration of phylogenetic effects and methodological issues about the spatial or temporal arrangement of behavioral tests did not improve model fit, implying that the effect of phylogeny remains undetectable even after controlling for other factors. Note that despite the weak phylogenetic effects, we detected strong roles for species-specific effects. Although phylogenetic inertia refers to constraints that act along the evolutionary history of species, species-specific roles depict the fact that different species demonstrate consistent variation in the strength of behavioral correlations. We found strong evidence for the latter but not for the former.

The large heterogeneity indicates that although the average effect size is significantly differentiable from zero, particular effects detected in different studies represent effects that have different biological meanings. We were able to identify some sources of this heterogeneity by modeling the effect of different moderators. Most importantly, some amount of the variation of effect sizes can be attributed to certain aspects of contextual overlap and species-specific roles. However, we cannot rule out the possibility that other factors also drive heterogeneity in effect sizes. We infer that population-specific effects or methodological differences among studies may be potential confounders that should be considered in future meta-analytic studies. The strong heterogeneity indicates that different selection pressures may shape different pair-wise associations differentially, also pointing to the main findings of our study.

CONCLUSIONS

Taken together, we found that although there was a general tendency for a positive phenotypic correlation among behavioral traits (Garamszegi, Markó, et al. 2012), the magnitude of these associations was generally low and varied considerably across pairwise correlations. We showed that spatial overlap among different behavioral tests could affect the detected strengths of certain syndromes, an issue that points to the importance of study design when studying behavioral syndromes. Furthermore, it seems that the commonly used behavioral variables in syndrome research (Réale et al. 2007) do not necessarily reflect equally different and independent biological domains. The detected correlations among particular traits might reflect varying ultimate and proximate mechanisms behind different behavioral syndromes as formed by different trait combinations. Results fail to provide support for the common organization of behaviors through a single individual background trait.

SUPPLEMENTARY MATERIAL

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

FUNDING

L.Z.G. was supported by the "Plan Nacional" program of the Spanish government (ref. no. CGL2009-10652 and CGL2009-09439). G.M. was supported by TÁMOP-4.2.1./B-09/1-KMR-2010-0005 and TÁMOP-4.2.2./B-10/1-2010-0023. G.H. received financial support from the Academy of Finland (128716), the Hungarian Scientific Research Fund (OTKA-K 105517) and János Bólyai Research Scholarship of the Hungarian Academy of Sciences.

An anonymous reviewer, Todd M. Freeberg, Shinichi Nakagawa, and Timothy H. Parker provided constructive comments during revision. Some authors provided useful explanations or unpublished information for calculating effect sizes, for which we are grateful (Peter Biro, Neeltje J. Boogert, Alecia J. Carter, Eleanor Cole, Niels Dingemanse, Sinéad English, Rebecca Fox, Todd M. Freeberg, Lauren Highfill, Katri Korpela, William James Loughry, Mark C. Mainwaring, Katherine A. Miller, Ariane Mutzel, Vladimir V. Pravosudov, John Quinn, Denis Réale, Alexandre Roulin, Heiko G. Rödel, Holger Schielzeth, Jessica Stapley, and Jana Uher). ManuSTAT Central offered statistical help.

Handling editor: Shinichi Nakagawa

REFERENCES

Adams DC. 2008. Phylogenetic meta-analysis. Evolution. 62:567–572.

Amy M, Sprau P, de Goede P, Naguib M. 2010. Effects of personality on territory defence in communication networks: a playback experiment with radio-tagged great tits. Proc R Soc B Biol Sci. 277:3685–3692.

Arnqvist G, Wooster D. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. Trends Ecol Evol. 10:236–240.

Barber I, Dingemanse NJ. 2010. Parasitism and the evolutionary ecology of animal personality. Phil Trans R Soc Lond B Biol Sci. 365:4077–4088.

Begg CB, Mazumdar M. 1994. Operating characteristics of a rank correlation test for publication bias. Biometrics. 50:1088–1101.

Bell AM. 2005. Behavioural differences between individuals and two populations of stickleback (*Gasterosteus aculeatus*). J Evol Biol. 18:464–473.

Bell AM, Hankison SJ, Laskowski KL. 2009. The repeatability of behaviour: a meta-analysis. Anim Behav. 77:771–783.

Bell AM, Sih A. 2007. Exposure to predation generates personality in three-spined sticklebacks (*Gasterosteus aculeatus*). Ecol Lett. 10:828–834.

Bell AM, Stamps JA. 2004. Development of behavioural differences between individuals and populations of sticklebacks, *Gasterosteus aculeatus*. Anim Behav. 68:1339–1348.

Blomberg S, Garland TJ, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 57:717–745.

Borenstein M. 2010. Comprehensive meta analysis—a computer program. Version 2.2.055. Englewood (NJ): Biostat Inc.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2009. Introduction to meta-analysis. West Sussex (UK): John Wiley & Sons, Ltd.

Brydges NM, Colegrave N, Heathcote RJP, Braithwaite VA. 2008. Habitat stability and predation pressure affect temperament behaviours in populations of three-spined sticklebacks. J Anim Ecol. 77:229–235.

Budaev SV. 1997. "Personality" in the guppy (*Poecilia reticulata*): a correlational study of exploratory behavior and social tendency. J Comp Psychol. 111:399–411.

Burghardt GM, Bartmess-LeVasseur JN, Browning SA, Morrison KE, Stec CL, Zachau CE, Freeberg TM. 2012. Minimizing observer bias in behavioral studies: a review and recommendations. Ethology. 118:511–517.

Buss AH, Chess S, Goldsmith HH, Hinde RA, McCall RB, Plomin R, Rothbart MK, Thomas A. 1987. What is temperament: four approaches. Child Dev. 58:505–529.

Carere C, Drent PJ, Privitera L, Koolhaas JM, Groothuis TGG. 2005. Personalities in great tits, *Parus major*: stability and consistency. Anim Behav. 70:795–805.

Carver CS, Connor-Smith J. 2010. Personality and coping. Annu Rev Psychol. 61:679–704.

Chamberlain SA, Hovick SM, Dibble CJ, Rasmussen NL, Van Allen BG, Maitner BS, Ahern JR, Bell-Dereske LP, Roy CL, Meza-Lopez M, et al. 2012. Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. Ecol Lett. 15:627–636.

Clark LA, Wilson D. 1999. Temperament: a new paradigm for trait psychology. In: Pervin LA, John OP, editors. Handbook of personality: theory and research. New York: The Guilford Press. p. 399–423.

Coats J, Poulin R, Nakagawa S. 2010. The consequences of parasitic infections for host behavioural correlations and repeatability. Behaviour. 147:367–382.

Cohen J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Erlbaum Associates.

Cohen J. 1992. A power primer. Psychol Bull. 112:155–159.

Congdon P. 2005. Bayesian predictive model comparison via parallel sampling. Comput Stat Data Anal. 48:735–753.

Conrad JL, Weinersmith KL, Brodin T, Saltz JB, Sih A. 2011. Behavioural syndromes in fishes: a review with implications for ecology and fisheries management. J Fish Biol. 78:395–435.

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. Control Clin Trials. 7:177–188.

Dingemanse NJ, Dochtermann NA, Nakagawa S. 2012. Defining behavioural syndromes and the role of "syndrome deviation" in understanding their evolution. Behav Ecol Sociobiol. 66:1543–1548.

Dingemanse NJ, de Goede P. 2004. The relation between dominance and exploratory behavior is context-dependent in wild great tits. Behav Ecol. 15:1023–1030.

Dingemanse NJ, Wright J, Kazem AJN, Thomas DK, Hickling R, Dawnay N. 2007. Behavioural syndromes differ predictably between 12 populations of three-spined stickleback. J Anim Ecol. 76:1128–1138.

Duval S, Tweedie R. 2000. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 56:455–463.

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika. 10:507–521.

Garamszegi LZ, Eens M, Török J. 2008. Birds reveal their personality when singing. PLoS One. 3:e2647.

Garamszegi LZ, Eens M, Török J. 2009. Behavioural syndromes and trappability in free-living collared flycatchers, *Ficedula albicollis*. Anim Behav. 77:803–812.

Garamszegi LZ, Markó G, Herczeg G. 2012. A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. Evol Ecol. 26:1213–1235.

Garamszegi LZ, Nunn CL, McCabe CM. 2012. Informatics approaches to develop dynamic meta-analyses. Evol Ecol. 26:1275–1276.

Garamszegi LZ, Rosivall B, Rettenbacher S, Markó G, Zsebők S, Szöllősi E, Eens M, Potti J, Török J. 2012. Corticosterone, avoidance of novelty, risk-taking and aggression in a wild bird: no evidence for pleiotropic effects. Ethology. 118:621–635.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. Stat Sci. 7:457–511.

Gelman A, Weakliem D. 2009. Of beauty, sex, and power. Am Sci. 97:310–316.

Gosling SD. 2001. From mice to men: what can we learn about personality from animal research? Psychol Bull. 127:45–86.

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

Herczeg G, Garamszegi LZ. 2012. Individual deviation from behavioural correlations: a simple approach to study the evolution of behavioural syndromes. Behav Ecol Sociobiol. 66:161–169.

Huntingford F, Coyle S. 2007. Antipredator defences in sticklebacks: tradeoffs, risk sensibility, and behavioural syndromes. In: Ostlund-Nilsson S, Mayer I, Huntingford FA, editors. Biology of the three-spined stickleback. Boca Raton (FL): CRC Press. p. 127–156.

Ioannidis JPA. 2005. Why most published research findings are false. PLOS Med. 2:e124.

Konishi S, Gupta AK. 1987. Inferences about interclass and intraclass correlations from familial data. In: MacNeil IB, Umphrey GJ, editors. Advances in the statistical sciences. Biostatistics. Vol. 4. Boston: Reidel. p. 225–233.

Lajeunesse MJ, Forbes MR. 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecol Lett. 6:448–454.

Lessells CM, Boag PT. 1987. Unrepeatable repeatabilities: a common mistake. Auk. 104:116–121.

Logue DM, Mishra S, McCaffrey D, Ball D, Cade WH. 2009. A behavioral syndrome linking courtship behavior toward males and females predicts reproductive success from a single mating in the hissing cockroach, *Gromphadorhina portentosa*. Behav Ecol. 20:781–788.

Mafli A, Wakamatsu K, Roulin A. 2011. Melanin-based coloration predicts aggressiveness and boldness in captive eastern Hermann's tortoises. Anim Behav. 81:859–863.

Maier SE, Vandenhoff P, Crowne DP. 1988. Multivariate analysis of putative measures of activity, exploration, emotionality, and spatial-behavior in the hooded rat (*Rattus norvegicus*). J Comp Psychol. 102:378–387.

Mateos-González F, Senar JC. 2012. Melanin-based trait predicts individual exploratory behaviour in siskins, *Carduelis spinus*. Anim Behav. 83:229–232.

McCrae RR, John OP. 1992. An introduction to the 5-factor model and its applications. J Pers. 60:175–215.

Møller AP, Jennions MD. 2002. How much variance can be explained by ecologists and evolutionary biologists. Oecologia. 132:492–500.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 82:591–605.

Nakagawa S, Santos ESA. 2012. Methodological issues and advances in biological meta-analysis. Evol Ecol. 26:1253–1274.

van Oers K, Drent PJ, de Goede P, van Noordwijk AJ. 2004. Realized heritability and repeatability of risk-taking behaviour in relation to avian personalities. Proc R Soc Lond Ser B Biol Sci. 271:65–73.

Quinn JL, Cresswell W. 2005. Personality, anti-predation behaviour and behavioural plasticity in the chaffinch *Fringilla coelebs*. Behaviour. 142:1377–1402.

R Development Core Team. 2007. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ. 2007. Integrating animal temperament within ecology and evolution. Biol Rev. 82:291–318.

Ruuskanen S, Laaksonen T. 2010. Yolk hormones have sex-specific longterm effects on behavior in the pied flycatcher (*Ficedula hypoleuca*). Horm Behav. 57:119–127.

Schaefer J, Opgen-Rhein R, Zuber V, Ahdesmaki M, Silva APD and Strimmer K. 2013. corpcor: efficient estimation of covariance and (partial) correlation. R package version 1.6.5. http://CRAN.R-project.org/package=corpcor.

Schielzeth H. 2010. Simple means to improve the interpretability of regression coefficients. Methods Ecol. Evol. 1:103–113.

Sih A, Bell AM. 2008. Insights for behavioral ecology from behavioral syndromes. Adv Study Behav. 38:227–281.

Sih A, Bell A, Johnson JC. 2004. Behavioral syndromes: an ecological and evolutionary overview. Trends Ecol Evol. 19:372–378.

Sih A, Bell AM, Johnson JC, Ziemba RE. 2004. Behavioral syndromes: an integrative overview. Q Rev Biol. 79:241–277.

Sih A, Cote J, Evans M, Fogarty S, Pruitt J. 2012. Ecological implications of behavioural syndromes. Ecol Lett. 15:278–289.

Sokal RR, Rohlf FJ. 1995. Biometry. 3rd ed. New York: W.H. Freeman & Co.

Spearman C. 1904. The proof and measurement of association between two things. Am J Physiol. 15:72–101.

Symonds MRE, Moussalli A. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol. 65:13–21.

Verbeek MEM, Boon A, Drent PJ. 1996. Exploration, aggressive behavior and dominance in pair-wise confrontations of juvenile male great tits. Behaviour. 133:945–963.

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. J Stat Softw. 36:1–48.

Walker DA. 2003. JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. J Mod Appl Stat Meth. 2:525–530.

Wilson ADM, Godin J-GJ. 2009. Boldness and behavioral syndromes in the bluegill sunfish, *Lepomis macrochirus*. Behav Ecol. 20:231–237.

Wilson DS. 1998. Adaptive individual differences within single populations. Phil Trans R Soc Lond B Biol Sci. 353:199–205.

Wingfield JC, Farner DS. 1993. Endocrinology of reproductive of wild species. In: Farner DS, King JR, Parkes KC, editors. Avian biology. New York: Academic Press. p. 163–327.

Wolf M, van Doorn GS, Leimar O, Weissing FJ. 2007. Life-history trade-offs favour the evolution of animal personalities. Nature. 447:581–584.

Wolf M, Weissing FJ. 2010. An explanatory framework for adaptive personality differences. Phil Trans R Soc Lond B Biol Sci. 365:3959–3968.